ehp ENVIRONMENTAL HEALTH PERSPECTIVES
http://www.ehponline.org

The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Integration of Animal and Human Evidence for PFOA Effects on Fetal Growth

Juleen Lam, Erica Koustas, Patrice Sutton, Paula I. Johnson, Dylan S. Atchley, Saunak Sen, Karen A. Robinson, Daniel A. Axelrad, and Tracey J. Woodruff

http://dx.doi.org/10.1289/ehp.1307923
Received: 22 November 2013
Accepted: 20 June 2014
Advance Publication: 25 June 2014

NIH National Institute of Environmental Health Sciences

The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Integration of Animal and Human Evidence for PFOA Effects on Fetal Growth

Juleen Lam,[1] Erica Koustas,[2] Patrice Sutton,[3] Paula I. Johnson,[3] Dylan S. Atchley,[3] Saunak Sen,[4] Karen A. Robinson,[5] Daniel A. Axelrad,[6] and Tracey J. Woodruff[3]

[1]Johns Hopkins University, Department of Health Policy & Management, Baltimore, Maryland, USA; [2]Oak Ridge Institute for Science and Education (ORISE) Post-doctoral Fellow with the U.S. Environmental Protection Agency, Office of Policy, National Center for Environmental Economics, Washington, DC, USA; [3]University of California San Francisco, Program on Reproductive Health and the Environment, Oakland, California, USA; [4]University of California San Francisco, Department of Epidemiology and Biostatistics, San Francisco, California, USA; [5]Johns Hopkins University, Departments of Medicine, Epidemiology and Health Policy and Management, Baltimore, Maryland, USA; [6]U.S. Environmental Protection Agency, Office of Policy, National Center for Environmental Economics, Washington, DC, USA

Address correspondence to Juleen Lam, Johns Hopkins University Bloomberg School of Public Health, 624 N. Broadway, Room 412, Baltimore, MD 21205 USA. Telephone: (443) 287-5324. E-mail: jlam5@jhu.edu

Running title: Integration of evidence for PFOA and fetal growth

Acknowledgments: The research was supported in part by appointments to the Internship/Research Participation Program at the National Center for Environmental Economics, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and

Education through an interagency agreement between the U.S. Department of Energy and EPA. Dr. Woodruff is also a member of the Philip R. Lee Institute for Health Policy Studies, UCSF. Dr. Lam is partially funded by the Environmental Defense Fund Washington DC. This research was also funded by the U.S. Environmental Protection Agency through a contract with Abt Associates (GAIA-0-6-UCSF 17288) and grants from New York Community Trust. Financial support over the past 5 years for the development and dissemination of the Navigation Guide methodology on which this case study was based was provided to UCSF Program on Reproductive Health and the Environment by: Clarence Heller Foundation, Passport Foundation, Forsythia Foundation, the Johnson Family Foundation, the Heinz Endowments, the Fred Gellert Foundation, Rose Foundation, Kaiser Permanente, New York Community Trust, University of California, San Francisco Institute for Health Policy Studies, Planned Parenthood Federation of America, National Institute for Environmental Health Sciences (ES018135 and ESO22841) and US Environmental Protection Agency STAR (RD83467801 and RD83543301).

We thank Kate Guyton (U.S. Environmental Protection Agency) for her contribution to developing the evidence evaluation methodologies and assistance with identifying relevant toxicological databases. We also thank Kristina Thayer (National Toxicology Program), Andy Rooney (National Toxicology Program), Lisa Bero (University of California, San Francisco), and Lauren Zeise (California Environmental Protection Agency) for assistance with developing risk of bias evaluation criteria, meta-analysis methodologies, and evidence evaluation methodologies. Andy Rooney also provided toxicological expertise and assisted with developing animal risk of bias criteria. Malcolm Macleod (University of Edinburgh) provided critical assistance in developing the animal-specific systematic review methodology. We thank Hanna Vesterinen (University of California, San Francisco) for her thoughtful comments on this

manuscript and Janet Pan (University of California, San Francisco) for assistance with data extraction. Christopher Lau (U.S. Environmental Protection Agency) served as a PFOA subject matter expert. We thank Tara Horvath (University of California, San Francisco) for training and assistance with developing our search strategy.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Competing financial interests: The authors declare no competing financial interests.

Abstract

Background: The Navigation Guide is a novel systematic review method to synthesize scientific evidence and reach strength-of-evidence conclusions for environmental health decision-making.

Objective: Integrate scientific findings from human and non-human studies to determine the overall strength of evidence for the question: "Does developmental exposure to perfluorooctanoic acid (PFOA) affect fetal growth in humans?"

Methods: We developed and applied *a priori* criteria to systematically and transparently: (1) rate the *quality* of the scientific evidence as 'high,' 'moderate' or 'low'; (2) rate the *strength* of the human and non-human evidence separately as: 'sufficient,' 'limited,' 'moderate,' or 'evidence of lack of toxicity'; and (3) *integrate* the strength of the human and non-human evidence ratings into a strength of the evidence conclusion.

Results: We identified 18 epidemiology and 21 animal toxicology studies relevant to our study question. We rated both the human and non-human mammalian evidence as 'moderate' quality and 'sufficient' strength. Integration of these evidence ratings produced a final strength of evidence rating where review authors concluded that PFOA is 'known to be toxic' to human reproduction and development based on sufficient evidence of decreased fetal growth in both human and non-human mammalian species.

Conclusion: The authors of this review concluded that developmental exposure to PFOA adversely affects human health based on sufficient evidence of decreased fetal growth in both human and non-human mammalian species. The results of this case study demonstrated the application of a systematic and transparent methodology, via the Navigation Guide, for reaching strength of evidence conclusions in environmental health.

Introduction

Evidence-based decision-making in environmental health requires synthesizing research from human and non-human (i.e., animal) evidence to reach overall strength of evidence conclusions, and is an integral part of hazard identification and risk assessment (National Research Council 2009). However, numerous shortcomings of current methods for research synthesis in environmental health have been identified—in particular, a robust, systematic and transparent methodology is needed (National Research Council 2011). To the extent that science informs decision-making, limitations in the methods for evaluating the strength of evidence in environmental health impedes our capability to act on the science in a timely way to improve health outcomes (Woodruff and Sutton 2014).

In the clinical sciences, methods of research synthesis—which integrate transparent and systematic approaches to evidence collection and evaluation—have been developed and refined over the past three decades and have played a transformative role in evidence-based decision-making for medical interventions (GRADE Working Group 2012; Higgins and Green 2011). For example, a systematic review and cumulative meta-analysis (continually updating the meta-analysis with results from more recent clinical trials) in cardiovascular medicine found discrepancies between recommendations by clinical experts and meta-analytic evidence. Experts often would not recommend treatments that pooled evidence demonstrated as effective, or even recommend treatments shown to have no effect or were potentially harmful (Antman et al. 1992). As a result, systematic and transparent methods of research synthesis are now relied upon in clinical medicine to determine which interventions should be offered to patients. Empirical evidence finds that this approach to evidence-based medicine is superior compared to traditional expert-based narrative reviews (Antman et al. 1992; Fox 2010; Rennie and Chalmers 2009).

However, methods of research synthesis used in the clinical sciences are not fully applicable to environmental health, primarily due to the differences in evidence streams and decision contexts between the two (Woodruff et al. 2011a). In particular, robust methods for evaluating non-human evidence streams, and fully developed methods for evaluating observational human studies are lacking (Woodruff et al. 2011a). In response to the need for improved methods of research synthesis in environmental health, beginning in 2009, an inter-disciplinary collaboration of 22 clinicians and scientists from federal and state government agencies, academic institutions, and non-governmental organizations developed the Navigation Guide systematic review method—see Supplemental Material, Navigation Guide Workgroup Members for additional details (Woodruff et al. 2011a). The Navigation Guide methodology incorporates best practices in research synthesis from clinical and environmental health science and provides an approach for evaluating and integrating human and non-human evidence streams (Woodruff et al. 2011a). The result of applying the Navigation Guide methodology is a concise statement about the quality and strength of the body of evidence of a contaminant's toxicity.

We undertook a case study to apply the Navigation Guide methodology and demonstrate the applicability of systematic and transparent methods of research synthesis to environmental health. In two systematic reviews we assessed human and non-human scientific evidence, including rating the quality and strength of evidence (Johnson et al. 2014; Koustas et al. 2014). In this paper, we integrated the strength of the human and non-human evidence ratings from these papers into an overall strength of evidence rating for an association between exposure to perfluorooctanoic acid (PFOA) and fetal growth. We selected this question because 1) PFOA has been in widespread use for over fifty years (Prevedouros et al. 2006; U.S. Environmental Protection Agency 2012); 2) PFOA is ubiquitous in the blood of the general U.S. population,

including pregnant women, women of child-bearing age, and in cord blood (Apelberg et al. 2007; Mondal et al. 2012; Woodruff et al. 2011b); 3) fetal growth is a health outcome of great public health importance (Institute of Medicine 2007); and 4) we were aware of multiple epidemiological and mammalian toxicology studies addressing this question available in the peer-reviewed scientific literature.

Methods

The Navigation Guide outlines four steps, the first three of which were addressed in this case study. *A priori*, we assembled a review team to include experts in the fields of risk assessment, environmental health, epidemiology, biology, systematic review, and toxicology to develop a protocol to address each step: 1) Specify the study question; 2) Select the evidence; and 3) Rate the quality and strength of the evidence (Woodruff et al. 2011a). The methods for each step were outlined *a priori* in protocols developed separately for human and non-human evidence (UCSF Program on Reproductive Health and the Environment 2013). The fourth and final step of the Navigation Guide, i.e., Grade Strength of Recommendation (to determine the final recommendation for public health protection), was not addressed in this case study due to resource constraints. Additional information regarding the Navigation Guide methodology and the review team can be found elsewhere (Woodruff and Sutton 2014).

Steps 1-3 are briefly summarized below for the human and non-human evidence streams. The detailed methods for each step in the human and non-human evidence streams are presented in separate papers (Johnson et al. 2014; Koustas et al. 2014). The present paper describes a novel feature of the Navigation Guide systematic review method, the process of integrating the quality and strength of the human and non-human bodies of evidence into a final strength of evidence conclusion about human toxicity.

Step 1. Specify the study question

Our overall objective was to integrate scientific findings from human and non-human studies to rate the strength of evidence for the question: "Does developmental exposure to perfluorooctanoic acid (PFOA) affect fetal growth in humans?" A "PECO" framework, which stands for "**P**opulation," "**E**xposure," "**C**omparator," and "**O**utcomes," was used to develop our question (Higgins and Green 2011). We established two separate PECO statements, one for human and one for non-human evidence (Table 1). These PECO statements were used to develop the search terms and inclusion/exclusion criteria for our systematic review in the next step.

Step 2. Select the evidence

We implemented a comprehensive search strategy to identify human and non-human studies from the scientific literature. We searched a variety of databases to identify studies, using search terms tailored for each database based on our PECO statements. We also hand searched the reference lists of included articles to identify additional studies. Our search was not limited by language or publication date.

All results were screened using *a priori* selection criteria using a structured form in DistillerSR (Evidence Partners; available at: http://www.systematic-review.net). Studies were excluded if one or more of the following criteria were met: article did not include original data (i.e., a review article); study did not evaluate humans or animals (i.e., *in vitro* studies); study subjects were not exposed to PFOA or exposure was not during the reproductive or developmental time period; or fetal growth or birth weight was not measured. From eligible studies, we collected details of the study characteristics, exposure assessment, outcome measurements and other information used to assess risk of bias using either a structured form in DistillerSR or Microsoft Access (2007). We

contacted study authors to request any data needed for the analysis that were not reported in the published articles.

Statistical analysis

For both human and non-human studies, we assessed study characteristics (i.e., study features and biological heterogeneity) to identify studies suitable for meta-analysis.

We used a random effects meta-analysis approach using the Der Simonian-Laird estimator of potential statistical heterogeneity across studies. All computations for the human studies meta-analysis were done in STATA version 12.1 (StataCorp LP, College Station, Texas, USA) using the "metaan" command. All computations for the non-human studies meta-analysis were done in the programming environment R version 2.13.1 (R Development Core Team; available at: http://www.R-project.org/), using the package "metafor" (Viechtbauer 2010).

In order to visually assess the possibility of publication bias in a meta-analysis, we considered producing a funnel plot of the estimated effects. However, tests for funnel plot asymmetry are not recommended when there are fewer than ten studies because test power is usually too low to distinguish chance from real asymmetry (Sterne et al. 2011a). As our meta-analysis for animals and humans was limited to less than ten studies each, we did not produce a funnel plot.

Statistical heterogeneity

We tested study variability using Cochran's Q statistic to detect whether differences in the estimated effect between studies could be explained by chance alone or due to non-random sources of variability between studies. We considered a p-value<0.05 to be statistically significant. We also calculated the I^2 statistic, which estimates the percent of variation across studies due to heterogeneity rather than chance (Higgins et al. 2003). To assess the impact of existing study heterogeneity on the meta-analysis, we considered the magnitude/direction of

effect estimates, the Cochrane Collaboration's guidelines to interpreting the I^2 values (Deeks et al. 2011), and statistical tests of heterogeneity (e.g., by assessing the p-value from the Cochran's Q test).

Sensitivity analysis

We conducted sensitivity analyses to investigate the effect on meta-analysis results. For the human evidence stream, we explored the effect of removing one included dataset at a time, as well as adding back in an excluded study. For the non-human evidence stream, we explored the effect of removing one included dataset at a time.

Step 3. Rate the quality and strength of the evidence

Figure 1 provides an overview of the rating process and includes risk of bias domains, quality of evidence factors, and strength of evidence considerations used to rate the quality and strength of the human and non-human evidence. We used this rating process to evaluate the human and non-human evidence streams separately.

Risk of bias across studies

Two review authors (JL and EK for non-human studies, PIJ and DSA for human studies) independently assessed each included study for the risk of bias, defined as study characteristics that may introduce a systematic error in the magnitude or direction of the results (Higgins and Green 2011). We developed an instrument for rating risk of bias by modifying existing risk of bias instruments used in human studies in the clinical sciences, i.e. the Cochrane Collaboration's Risk of Bias tool and the Agency for Healthcare Research and Quality's (AHRQ) criteria (Higgins and Green 2011; Viswanathan et al. 2012).

The Cochrane Collaboration's Risk of Bias Tool does not currently include a specific domain for bias related to study funding source, but this is an area of active discussion among its members (Bero 2013; Sterne 2013). The Collaboration has recognized the importance of identifying study funding source, which has been empirically shown to be associated with biases (Krauth et al. 2014; Lundh et al. 2012). However, there is currently limited consensus on whether study funding source should be included as a separate risk of bias domain or generally reported and commented on within the Cochrane systematic review (Bero 2013; Sterne 2013). A recent report from the National Research Council (NRC) recommended that the U.S. Environmental Protection Agency (EPA) consider funding sources in their risk of bias assessment conducted for systematic reviews (National Research Council 2014).

Therefore, we also included study funding source and declared financial conflicts of interest as a potential source of bias (i.e., whether the study received support from a company, study author, or other entity having a financial interest in the outcome of the study). We refer to this risk of bias domain generally as "Conflicts of interest", although for this particular case study we only assessed competing financial interests within this domain. See Figure 1 for a complete list of the human- and non-human risk of bias domains; detailed descriptions of each domain are available elsewhere (Koustas et al. 2014; Johnson et al. 2014). Each risk of bias domain was assigned a determination of *high*, *probably high*, *low*, or *probably low* risk of bias based on *a priori* determined criteria (Koustas et al. 2014; Johnson et al. 2014). We followed the Grading of Recommendations Assessment Development and Evaluation (GRADE) principles for evaluating overall risk of bias by judiciously considering the frequency of each type of bias across all studies, evaluating the extent to which each study contributed toward the magnitude of effect estimate, and being conservative in the judgment of rating down (i.e., evidence was only rated

down if risk of bias was clearly a substantial issue across most studies) (Viswanathan et al. 2012).

Rating the quality of evidence across studies

Each of the review authors compared the results of the systematic review to the Navigation Guide factors and considerations for rating the quality of the evidence as a way to initiate the group discussion and gather all perspectives for consideration. The Navigation Guide rating method (Woodruff and Sutton 2011) was applied according to explicit written directions (Koustas et al. 2014; Johnson et al. 2014). The possible ratings for the overall quality of evidence were 'high,' 'moderate,' or 'low.' Adapting the GRADE method as guidance, we first assigned an *a priori* initial quality rating to the body of evidence, and then considered adjustments ("downgrades" or "upgrades") to the quality rating based on the characteristics of the studies constituting the body of evidence to arrive at a final rating determination (Balshem et al. 2011).

We assigned *a priori* initial ratings of 'moderate' for the body of human observational data and 'high' for the experimental non-human data. We assigned the body of human observational studies an initial rating of 'moderate' independent of the specifics of included studies; these characteristics were then evaluated later for upgrading or downgrading this rating. Our rationale to assign the initial rating of 'moderate' was based on the absolute and relative merit of human observational data in evidence-based decision-making in environmental and clinical sciences. Human observational studies generally are recognized as being a reliable source of evidence in the clinical sciences and the preferred method for evaluating disease etiology (Institute of Medicine et al. 2008). As ethical considerations virtually preclude experimental human data from the environmental health evidence stream, human observational studies are typically the "gold standard" of this evidence base. In comparison, randomized animal experiments have a high

level of study design control, including level and duration of exposure, and test a study population of limited heterogeneity (inbred strains of laboratory animals). Thus, these data were the most comparable to human randomized controlled trials (RCTs) in the clinical sciences and therefore we assigned the experimental non-human data (both mammalian and non-mammalian) the initial rating of 'high' to reflect this.

We assessed the overall body of human evidence for downgrading and upgrading the *a priori* 'moderate' quality rating based on 8 factors—5 for downgrading and 3 for upgrading. Our criteria for evaluating evidence from studies incorporate elements similar to the Bradford-Hill considerations (i.e., consistency of effect, strength of effect, biologic gradient as well as incorporating experimental evidence from animal studies) and elements from other frameworks for evaluating scientific evidence (from the U.S. Preventative Service Task Force and IARC) (International Agency for Research on Cancer 2006; Sawaya et al. 2007).

We decided to evaluate the non-human evidence separately for mammalian versus non-mammalian evidence due to fundamental biological differences between the two and the lower quality, i.e., high risk of bias, of the non-mammalian evidence. We evaluated each using the same 5 factors for downgrading the *a priori* 'high' quality rating, but did not consider any upgrades to the quality rating because the initial rating was already set at 'high.' Consistent with GRADE guidelines (Guyatt et al. 2011c), we did not upgrade or downgrade the body of evidence unless there was compelling rationale to do so.

Each of the nine review authors applied their expert judgment to review the bodies of evidence and independently graded the quality of evidence based on the presence of these factors using detailed instructions. The instructions to review authors contained specific information on how to

evaluate the quality of evidence; see Supplemental Material, Instructions for Rating the Quality and Strength of Human and Non-Human Evidence which are also available online (UCSF Program on Reproductive Health and the Environment 2013). Possible ratings were 0 (no change), -1 (1 level downgrade) or – 2 (2 level downgrade). Each overall body of evidence was evaluated for downgrading based on consideration of five factors:

1. Risk of bias across studies: Evidence streams were rated down if most of the relevant evidence came from studies that had high risk of bias, although review authors were instructed to be conservative in the judgment of rating down. In other words, review authors were instructed to rate down only if they judged there to be a substantial risk of bias in the body of available evidence. Furthermore, review authors were instructed not to assess factors by averaging across studies (e.g., if some studies had low risk of bias, a similar number of studies had probably high risk of bias, and a similar number of studies had high risk of bias, the quality should not be downgraded solely by averaging the risk of bias ratings).
2. Indirectness**:** Following GRADE guidelines (Guyatt et al. 2011a), evidence streams were rated down if substantial differences existed between the study population, exposure, comparator, or outcome measured as compared to those for our study question. Potential sources of indirectness included if the study population or intervention/exposure was so different from that of interest that there was a compelling reason to think that the magnitude of effect would differ substantially, or if studies reported on surrogate endpoints in place of the outcome of interest. In contrast to GRADE, our *a priori* assumption is that animal evidence is direct evidence of human health. However, in applying GRADE principles to the Navigation Guide, animal evidence will be rated

down if it is determined that it is a biologically inappropriate non-human model for the health outcome under study.

3. Inconsistency: Evidence streams were rated down if studies had widely different estimates of effect (unexplained heterogeneity or variability in results) looking across studies conducted in similar human populations or animal species. The following considerations were used to indicate potential "inconsistency": if point estimates varied widely across studies, confidence intervals showed minimal or no overlap for similar studies of comparable size, the statistical test for heterogeneity showed a low p-value ($p<0.05$); and/or the I^2 was large (>50%, based on the Cochrane's guide to interpretation of I^2) (Higgins and Green 2011)). Review authors were instructed to downgrade only when the inconsistent findings reduced confidence in the results in relation to the direction of effect estimates (i.e., studies that were inconsistent with respect to the magnitude of an effect (but not in terms of direction of effect estimates) would not be rated down).
4. Imprecision: Evidence streams were rated down if most studies had small sample sizes and few events, thus leading to wide confidence intervals.
5. Publication bias: Evidence steams were rated down if we thought that studies were missing from the body of evidence that might result in an overestimate or underestimate of true exposure effects. We used considerations from GRADE guidance for evaluating publication bias, with modifications to reflect the Navigation Guide's primary concern with underestimating the true effects of existing chemical exposure, in contrast to GRADE's primary concern of overestimating the true effect of treatments or pharmaceuticals (Guyatt et al. 2011d). These modified considerations included: if the

body of evidence was dominated by early studies with negative results, particularly if they were small in size; studies were uniformly small (particularly when sponsored or funded by industry); empirical examination of patterns of results (e.g. funnel plots) suggest publication bias; there was success in obtaining results of unpublished studies that demonstrated different results from published studies; and/or a comprehensive search of the literature was not performed.

Furthermore, the rating of each factor was considered in the context of other limitations. For instance, if review authors found themselves in a close-call situation with respect to two quality issues (i.e., "risk of bias across studies" and "imprecision"), we followed the suggestion from GRADE to rate down for at least one of the two factors (Guyatt et al. 2011c).

The instructions to review authors also contained information on how to evaluate the human body of evidence for upgrading based on consideration of three factors (animal evidence was not eligible for upgrading since it started at an initial 'high' rating); see Supplemental Material, Instructions for Rating the Quality and Strength of Human and Non-Human Evidence. Possible ratings were 0 (no change), +1 (1 level upgrade) or +2 (2 level upgrade):

1. Large magnitude of effect: Recommendations from the GRADE group (Guyatt et al. 2011c) are to rate the evidence stream up by one category (for instance, from 'low' to 'moderate') if there were associations with a relative risk (RR) greater than 2, and up by two categories (for instance, from 'low' to 'high') for those with RR greater than 5. However, there are limitations to using RR to determine magnitude of effect, as it relies on dichotomous exposure scales and outcomes. Although there is no established cutoff for the continuous scales, we evaluated the evidence judiciously to assess whether the

magnitude of effect from the human evidence was compelling enough to justify upgrading the evidence.

2. Dose-response gradient: The evidence stream was rated up if there were consistent dose-response gradients within one or multiple studies, and/or evidence of dose-response across the studies in the overall body of evidence.
3. Confounders minimize the demonstrated effect: The evidence stream was rated up if consideration of plausible residual confounders or biases would only reduce the magnitude of observed effect, or suggest a spurious effect when results show no effect. GRADE provides an illustrative example of rating up observational evidence finding lack of association between vaccination and autism, which occurred despite empirically confirmed bias that parents of autistic children may be more likely to remember their vaccine experience. The negative findings despite this form of recall bias suggest rating up the quality of evidence (Guyatt et al. 2011c).

Consistent with GRADE's approach to evaluating risk of bias across studies (Guyatt et al. 2011), authors were instructed to be conservative in making judgments to downgrade the evidence for all factors (i.e. high confidence in concerns with the body of evidence before rating down). After independently evaluating the quality of the evidence, all authors collectively discussed their evaluations. This discussion between co-authors was extensive, iterative, and carried out over several meetings until a consensus was reached. Specifically, these collective decisions did not involve a "majority vote" or other tallying of perspectives. It was specified *a priori* that discrepancies between the review authors' judgments that could not be resolved through consensus would be resolved by the senior author (TW). However, for this case study review authors were able to agree on a collective consensus for each rating and the arbiter was not

necessary. The collective rationale for each decision on each of the factors was documented and recorded.

Rating the strength of the evidence across studies

In systematic reviews in the clinical sciences, only human evidence is considered in a decision, and so there exists no corollary step for integrating multiple streams of evidence in Cochrane or other methods of systematic review in the clinical sciences. We followed guidance from the International Agency for Research on Cancer (IARC) and toxicity definitions used by the U.S. EPA to develop our approach to rate the strength of evidence for the human and non-human bodies of evidence (International Agency for Research on Cancer 2006; U.S. Environmental Protection Agency 1991, 1996; Rooney et al. 2014).

We rated the overall strength of the human and non-human evidence separately based on a combination of four considerations, which were developed from existing criteria for evaluating evidence streams (International Agency for Research on Cancer 2006): (1) Quality of body of evidence (i.e., our rating from the previous step); (2) Direction of effect estimates; (3) Confidence in effect estimates (likelihood that a new study would change our conclusion); and (4) Other compelling attributes of the data that may influence certainty (Figure 1). We compared the results of rating the strength of the human and non-human evidence to the definitions specified in the Navigation Guide for 'sufficient evidence of toxicity,' 'limited evidence of toxicity,' 'inadequate evidence of toxicity,' or 'evidence of lack of toxicity' to select one of these final ratings for each body of evidence. Detailed definitions for each rating can be found elsewhere (Johnson et al. 2014; Koustas et al. 2014).

Review authors independently evaluated the strength of the evidence according to the four considerations specified above to form their opinion of the final rating of strength of evidence for the human and non-human evidence as a way to initiate the group discussion and gather all perspectives for consideration. All authors collectively discussed their evaluations in a meeting until a consensus was reached. Specifically, this final rating did not involve a "majority vote" or other tallying of perspectives. It was specified *a priori* that discrepancies between the review authors' judgments that could not be resolved through consensus would be resolved by the senior author (TW). However, for this case study review authors were able to agree on a collective consensus for the final rating for strength of evidence and the arbiter was not necessary. The rationale for our collective decision on each of the criteria and overall ratings was documented and recorded.

Integration of the strength of human and non-human streams of evidence

The final step of our review was to integrate the strength of the human and non-human streams of evidence into a final concluding statement about PFOA toxicity. We compared the strength of the human and non-human evidence ratings to the integration table in Step 3 of the Navigation Guide, which was based on the method used by IARC and used their descriptors of strength of evidence, modified to be relevant for non-carcinogenic assessments (International Agency for Research on Cancer 2006; Woodruff and Sutton 2014).

By determining the intersection on this table of the ratings assigned to the human evidence (listed in the rows of the integration table in Step 3) and non-human evidence (columns of the integration table) (Woodruff and Sutton 2014), we came to one of the five possible strength of evidence conclusions about toxicity: 'known to be toxic,' 'probably toxic,' 'possibly toxic,' 'not classifiable,' or 'probably not toxic.' Importantly, consistent with IARC's strength of evidence

conclusions for cancer endpoints, 'sufficient evidence of toxicity' in humans would result in a 'known to be toxic' final conclusion, regardless of the non-human evidence rating. However, 'limited evidence of toxicity' in humans could result in a 'probably toxic' final conclusion if there was 'sufficient evidence of toxicity' in animals or a 'possibly toxic' final conclusion if there were 'limited,' 'inadequate,' or 'evidence of lack of toxicity' in animals. The terminology for these conclusions were adapted from IARC's methods for integrating human and non-human evidence (International Agency for Research on Cancer 2006), which in turn were linked to strength of evidence descriptions in use by U.S. EPA (U.S. Environmental Protection Agency 1991, 1996).

Results

Included studies

Our database and hand searching of human literature retrieved a total of 3,024 unique records—of these, we identified a total of 18 relevant studies (which contributed 19 datasets) for analysis (Figure 2). Our database and hand searching of the non-human literature retrieved a total of 2,049 unique records—of these we identified a total of 21 relevant studies (which contributed 32 relevant datasets) for analysis (Figure 2). There were more datasets than studies for both human and non-human evidence because some studies contributed multiple datasets, for example, if they measured several relevant outcomes or reported outcomes for different species or populations.

Risk of bias assessment

A summary of the risk of bias determinations is shown elsewhere (Johnson et al. 2014; Koustas et al. 2014). Potential sources of risk of bias occurring frequently in human studies were confounding, exposure assessment and conflict of interest. Potential sources of risk of bias

occurring frequently in non-human studies were inadequate sequence generation, allocation concealment, and blinding.

Statistical analysis

We combined data from 9 human studies in a meta-analysis of the effect of PFOA exposure on birth weight. The studies not included in the meta-analysis were determined to be not combinable with the others due to differences in PFOA exposure scale or outcome statistic (Arbuckle et al. 2012; Halldorsson et al. 2012; SK Kim et al. 2011; Monroy et al. 2008; Nolan et al. 2009; Savitz et al. 2012; Stein et al. 2009). We found from the meta-analysis an overall estimate of -18.9 grams birth weight per ng/mL increase in serum PFOA (95% confidence interval: -29.8, -7.9) (Johnson et al. 2014). The I^2 was 38%, indicating little heterogeneity between studies that could not be explained by chance; this was further supported by the Q statistic (p-value=0.12). Additional meta-analyses demonstrated that PFOA exposure was also slightly associated with decreases in other fetal growth measures at birth, such as length (n=5, overall estimate -0.1, 95% confidence interval: -0.1, -0.02), ponderal index (n=4, overall estimate -0.01, 95% confidence interval: -0.03, 0.01), and head circumference (n=4, overall estimate -0.03, 95% confidence interval: -0.1, 0.01) (Johnson et al. 2014).

Fifteen of the 21 non-human studies were conducted on mammalian species (11 mouse and 4 rat) and 6 were conducted on non-mammalian species (3 chicken, 1 fruit fly, 1 zebrafish, and 1 salmon) (Koustas et al. 2014). From an assessment of *a priori* determined considerations regarding study characteristics (e.g., species, route of exposure, method of outcome measurement, and time point of outcome measurement), we determined that seven of these studies (eight datasets) which all exposed pregnant mice through gavage PFOA treatments and measured weight of offspring at or soon after birth were suitable for meta-analysis.

We used the mean pup body weight at birth (and standard error) from each of the datasets, for all doses below 5 mg/kg-bw/day. The dose was limited to focus on effects at lower tested doses and minimize adverse impacts from responses at higher doses (such as litter loss) on the overall estimate. We initially attempted to transform animal tested doses to human-equivalent serum concentrations for more direct comparisons to the human data; however, a review of the available scientific data produced minimal data that would support such extrapolation. Review authors felt that our limitation to doses below 5 mg/kg-BW/day was adequate to ensure relevance of the animal dose-response estimates to humans. Furthermore, by using the slope of the dose-response model for animals our interpretation makes the assumption that similar increases in exposure would result in the same relative changes in birth weights compared to humans, which review authors considered reasonable. We found from the meta-analysis an overall estimate of -0.023 grams birth weight per mg/kg BW/day increase in PFOA dose to pregnant dams (95% CI: -0.029, -0.016) (Koustas et al. 2014). The I^2 test statistic was 0%, indicating no observed statistical heterogeneity between studies that could not be explained by chance; this conclusion was further supported by the Q statistic (p-value=0.73).

We also visually inspected scatter plots of dose-response data for all mammalian and non-mammalian animal data, including data excluded from the meta-analysis (those with study characteristic determined to be too variable to combine) to investigate effects (Koustas et al. 2014). The dose-response data from the eight mammalian datasets included in the meta-analysis showed similar results in the same direction (decreased weight) with mostly statistically significant results. In contrast, the dose-response data for the nine mammalian studies not included in the meta-analysis showed mixed results, generally with lower doses showing increased weight compared to the control group (mostly non-significant) and higher doses

showing decreased weight (both statistically significant and not). A qualitative evaluation of data for the non-mammalian studies showed mostly non-statistically significant increases in body weight (seen in multiple chicken studies, but not in fruit fly or salmon studies although there was only one study in each with a small number of tested doses). The length data for non-mammalian studies showed mixed results, including statistically significant decreases in length (in fruit flies and zebrafish) and the other two studies showing insignificant increases in length (in chickens and salmon); these discrepancies in part justify our decision to rate the body of non-mammalian studies overall to be of "low" quality (Koustas et al. 2014).

Sensitivity analysis of human studies demonstrated little change in the overall effect estimate when removing one included study at a time or adding in one excluded study, although the heterogeneity statistics did increase. Sensitivity analysis of the non-human studies when removing one included study at a time demonstrated little change in the overall effect estimate or heterogeneity statistic. We originally planned to produce funnel plots of the estimated effects to visually assess the possibility of publication bias, but we did not due to the small number of included studies.

Quality of the body of evidence

We evaluated each of the five quality downgrade factors separately for human, non-human mammalian, and non-mammalian streams of evidence. We concluded there was no indication of substantial "risk of bias across studies" for the available human evidence, particularly when evaluating the studies included in meta-analysis, so we did not downgrade the human evidence for this factor. The majority of non-human mammalian studies had *probably high* risk of bias for the "allocation concealment" and "blinding" domains. The non-mammalian studies had *probably high* risk of bias for the "sequence generation", "allocation concealment", and "blinding"

domains. Since these components have been shown empirically to influence study outcomes in experimental animal studies (Bebarta et al. 2003; Landis et al. 2012; Macleod et al. 2004), our group consensus was to downgrade each non-human body of evidence by one quality level (-1) for "risk of bias across studies."

We concluded there was no indication of substantial "indirectness" in either the body of available human or non-human mammalian evidence. The human studies assessed the population, exposure and outcomes of interest, as did the non-human mammalian evidence, based on empirical evidence that mammalian data can be used as direct evidence for human health inference (Kimmel et al. 1984; U.S. Environmental Protection Agency 1996). However, we could not identify a rationale or empirical basis for assuming directness of the non-mammalian body of evidence reviewed in this case study, and in particular, we were concerned about indirectness of the route of exposure (e.g., injection or immersion of eggs in PFOA-containing solution) and developmental differences (*in utero* development versus egg development) between humans and the non-mammalian model systems. Therefore, we downgraded the non-mammalian evidence one quality level (-1) for "indirectness."

We concluded there was no indication of "inconsistency" in any of the three bodies of evidence. With the exception of two small studies (Fromme et al. 2010; S Kim et al. 2011), results across the human studies were generally consistent in the magnitude and direction of effect estimates. This was further supported by the consistency of the overall meta-analysis results, which were minimally affected by results of any individual study, as determined by sensitivity analysis. For non-human mammalian studies, point estimates were generally consistent with overlapping confidence bounds, and meta-analysis results were consistent in direction of effect estimates and minimally affected by the results of any individual study, as determined by sensitivity analysis.

Non-mammalian studies differed based on outcome of measurement (weight vs. length), but results were consistent between comparable studies (similar outcome, species, and exposure route). Therefore, we did not downgrade the quality level for any of the bodies of evidence for "inconsistency."

We concluded there was no indication of "imprecision" in any of the three bodies of evidence. We judged the confidence intervals for both the human and non-human mammalian meta-analysis to be sufficiently narrow so as not to warrant downgrading the evidence. Similarly, confidence intervals for the non-mammalian evidence were either sufficiently narrow, or if none were given, the data showed statistically significant responses at high doses, indicating small confidence bounds. The group consensus after evaluating this factor was to not downgrade the quality level for any of the bodies of evidence for "imprecision."

We concluded there was no indication of "publication bias" in any of the three bodies of evidence. The literature search was comprehensive and included strategies to search the grey literature, such as conference abstracts, reports or other non peer-reviewed literature. Although we could not ensure we had identified all unpublished studies, the studies we found had varying sample sizes and funding sources, and no unpublished studies were found that presented results out of the range of estimates reported by published studies. Without a sufficient number of studies to produce an informative funnel plot to derive evidence about potential missing data, the group consensus was that we did not have substantial evidence to warrant downgrading the quality level for any of the bodies of evidence for "publication bias."

We evaluated each of the three upgrade quality factors for human evidence only. We found no compelling evidence to warrant upgrading the evidence based on our *a priori* definitions for the

three considered factors. We evaluated the human effect estimates judiciously and agreed that the magnitudes of effect estimates were not compelling enough to justify upgrading the evidence. Although several studies showed some evidence of a dose-response relationship, we agreed that the evidence was not compelling enough across the body of evidence as a whole. We also agreed that there was no evidence to suggest that consideration of plausible residual confounders or biases would reduce the estimated effect. The group consensus after evaluating these factors was to not upgrade the quality level for the human evidence.

A summary of our final decisions for each upgrade/downgrade factor for each of the three bodies of evidence is shown in Table 2. An assessment of these decisions resulted in an overall quality of the human evidence rating of 'moderate.' The overall quality rating of the non-human mammalian evidence was downgraded from 'high' to 'moderate' based on the "risk of bias across studies" factor. The overall quality rating of the non-mammalian evidence was downgraded from 'high' to 'low' based on the concerns regarding both the "risk of bias across studies" and "indirectness" factor.

Strength of the body of evidence rating

We rated the overall strength of the human and non-human bodies of evidence separately based on the four considerations: (1) Quality of body of evidence; (2) Direction of effect estimates; (3) Confidence in effect estimates (likelihood that a new study would change our conclusion); and (4) Other compelling attributes of the data that may influence certainty. Because the non-mammalian evidence quality was rated 'low' whereas the non-human mammalian data were 'moderate,' we made the decision to only carry forth the higher quality non-human mammalian body of evidence for evaluating strength of evidence. This is consistent with GRADE

recommendations: when high quality data are available for decision-making, one does not need to incorporate low quality data (Balshem et al. 2011).

We rated the "quality of body of evidence" for both human and non-human evidence as 'moderate,' as discussed in the previous section. The "direction of effect estimates" for both human and non-human evidence was assessed by evaluating across individual studies available as well as using results from the meta-analyses. We concluded that there was similar evidence of an association between decreased birth weight and increased exposures to PFOA for both evidence streams.

We evaluated the "confidence in effect estimates" using slightly different approaches for each body of evidence. For the human evidence, we used an ad hoc approach of quantitatively evaluating the potential impact of adding a new hypothetical study on the overall meta-analysis result. We considered several scenarios of adding a hypothetical study with characteristics similar to our included human studies to determine what effect estimates would be needed to alter the interpretation of our final meta-analysis result. Comparing this to the effect estimates of our included human studies, we decided that it seemed unlikely that another human study would find such associations. More details, including the quantitative estimates, may be found elsewhere (Johnson et al. 2014). For the non-human evidence, we determined that our confidence in the effect estimates was high because the results among non-human mammalian experimental studies were similar and demonstrated overlapping confidence intervals across different studies (Koustas et al. 2014). Lastly, we did not identify any other compelling attributes of the data that would influence our certainty in the estimates. In particular, we considered a hypothesis proposed in the literature whereby women who have smaller babies have higher measures of PFOA due to a lower glomerular filtration rate caused by lower plasma volume expansion. As

discussed below, we evaluated the supporting scientific evidence for this hypothesis in the context of our final conclusion from this review, and decided that it did not undermine our findings for several reasons.

A summary of our strength of evidence determinations for each consideration for human and non-human evidence is shown elsewhere (Johnson et al. 2014, Koustas et al. 2014). We compared these determinations to the definitions to evaluate the overall strength of each body of evidence (Johnson et al. 2014, Koustas et al. 2014). Our consensus for the human evidence was that the overall quality of evidence was 'moderate' and we had a high level of confidence in an association between decreased birth weight and increased exposures to PFOA. Comparing our consensus on these considerations to the definitions of 'sufficient evidence of toxicity,' 'limited evidence of toxicity,' 'inadequate evidence of toxicity,' or 'evidence of lack of toxicity,' we agreed our findings met the definitions for 'sufficient evidence of toxicity', i.e., a positive relationship was observed between exposure and outcome where chance, bias, and confounding could be ruled out with reasonable confidence; the available evidence included results from one or more well-designed, well-conducted studies; and the conclusion was unlikely to be strongly affected by the results of future studies.

Our consensus for the non-human studies was that the overall body of evidence was 'moderate' and we had a high level of confidence in an association between decreased birth weight and increased exposures to PFOA. We agreed our findings for the non-human (mammalian) studies met the definitions for 'sufficient evidence of toxicity', i.e., a positive relationship was observed between exposure and adverse outcome in multiple studies or a single appropriate study in a single species; the available evidence included results from one or more well-designed, well-

conducted studies; and the conclusion was unlikely to be strongly affected by the results of future studies

Our final conclusion for the overall strength of evidence was that there was 'sufficient evidence of toxicity' in humans and 'sufficient evidence of toxicity' in non-human mammalians to support a judgment that exposure to PFOA affects fetal growth.

Integrating the evidence across evidence streams

We integrated our evidence rating of 'sufficient evidence of toxicity' for the human and the non-human evidence and concluded that PFOA should be classified as 'known to be toxic.'

Discussion

The application of the Navigation Guide systematic review methodology demonstrated a novel method for integrating diverse sources of toxicity data to reach strength of evidence conclusions for non-cancer health effects in environmental health. Application of the method produced a clear and concise conclusion by the authors of this review: that "exposure to PFOA is 'known to be toxic' to human reproduction and development based on sufficient evidence of decreased fetal growth in both human and non-human mammalian species."

We concluded for the human data that there was 'sufficient' evidence of an association based on a transparent collective rating of the evidence as 'moderate' quality, a meta-analysis estimating a reduction in birth weight in relation to PFOA exposure in which confidence bounds were judged to be sufficiently narrow and did not include zero, and our confidence that it would be unlikely for a new study to have an effect estimate that could substantially change the overall effect estimate of the meta-analysis (Johnson et al. 2014). Similarly, we concluded for the non-human data that there was 'sufficient' evidence of an association based on a transparent collective rating

of the available non-human mammalian evidence as 'moderate' quality, a meta-analysis showing a reduction in birth weight in relation to PFOA dose with in which confidence bounds were judged to be sufficiently narrow and did not include zero, and our confidence that the conclusion was unlikely to be strongly affected by the results of future studies (Koustas et al. 2014).

In applying the Navigation Guide methodology to this case study, we found that the definitions used to rate the quality and strength of the evidence drive the final strength of evidence statement. While the domains and factors used for rating quality of evidence were derived from methods applied in the clinical sciences (Guyatt et al. 2008; Higgins and Green 2011), there is no precedent for defining and integrating strength of evidence conclusions among different evidence streams in the clinical sciences. Our definitions for strength of evidence and the process for integration of the evidence streams were derived from current practices in use by IARC (International Agency for Research on Cancer 2006) and EPA (U.S. Environmental Protection Agency 1991, 1996). Notably, while the Navigation Guide currently requires 'sufficient' human evidence for a chemical to be rated as 'known to be toxic,' this requirement may be revised in future case studies to align with other established methods in environmental health in which this requirement is not necessary (International Agency for Research on Cancer 2006; U.S. Environmental Protection Agency 1991, 1996; Rooney 2014). Given that the risk of bias criteria, the factors used to rate quality across a body of evidence, and the considerations for rating strength of evidence underlie the final integration step, research to deepen our knowledge of the relative and absolute impact of each of these criteria, factors and considerations in the final strength of evidence rating is currently a critical need.

We found that specific *a priori* definitions made rating the evidence at hand efficient and transparent. First, establishing precise *a priori* definitions ensured that we were all using the

same rules to apply our judgment and ensured our collective decisions were transparent and explicit even to ourselves. Second, determining definitions *a priori* encouraged us to actively think through the sources of data and the evidence necessary to support different conclusions regarding weight/strength of evidence, and identify how to establish scientifically valid definitions. While the definitions we used in this first case study can guide development of definitions for future case studies, they are not rigid and can potentially be refined to apply to any particular question and available body of evidence at hand.

Although the protocol defined many of the guidelines for making decisions *a priori*, we found we could not anticipate all decision-points beforehand. For example, we did not anticipate our search would retrieve such a diversity of non-mammalian model systems data (such as zebrafish and chickens) and we had to interpret the heterogeneity and relevance of these data to human health during the analysis. In another example, in following recommendations from GRADE, we defined the factor for upgrading the quality of evidence based on large magnitude of effect as associations with a relative risk (RR) greater than 2 (+1 upgrade to the evidence) or a RR greater than 5 (+2 upgrade to the evidence). However, the data from the human evidence were more amenable to a meta-analysis done on a continuous scale and therefore we did not have RRs to compare using this definition. Furthermore, RRs on a scale of 2 or 5 for non-occupational studies are a rarity in the field of environmental health, due to the relatively low levels of exposure to environmental contaminants (Taubes 1995). Therefore, although this is an accepted cutoff generally for GRADE, the definition of large magnitude of effect will require adjustment based on the nature and extent of the available evidence. This also may require additional consideration because the size of RR estimates is dependent upon the study author's selection of the

comparator group. Therefore, the definition of large magnitude of effect may also require adjustment based on the design of included studies and the specific biological outcomes.

On a similar note, for this case study we decided *a priori* to define the "inconsistency" factor to rate down each body of evidence if studies show widely different estimates of effect, but did not include a "consistency" factor to rate up each body of evidence for the converse scenario. This was done to ensure all bodies of evidence would be evaluated for consistency—since the non-human evidence was not assessed for upgrade factors because it started at 'high', we included "inconsistency" as a downgrade. This is consistent with GRADE recommendations for evaluating "inconsistency" for human evidence (Guyatt et al. 2011b). Again, this determination is not rigid and can be adjusted for future case studies. As an example of this, the recent proposal by the National Toxicology Program's (NTP) Office of Health Assessment and Translation (OHAT) for systematic review and evidence integration for health assessments instead includes "consistency" as a factor that increases confidence in the body of evidence, as opposed to our "inconsistency" downgrade factor (National Toxicology Program 2013). The approach to categorizing these factors may change, but the underlying consistency and transparency of each approach to evaluate the bodies of evidence is what is most important.

In recent years, several scientists have hypothesized that maternal and fetal physiology may influence measured blood levels of an exposure, and in particular for PFOA and reduced birth weight these associations may be due to reverse causality whereby women who have smaller babies have higher measures of PFOA due to a lower glomerular filtration rate caused by lower plasma volume expansion (Loccisano et al. 2013; Savitz 2007; Whitworth et al. 2012). If this reverse causality hypothesis were true, it could explain some or all of the relationship observed

in human cross-sectional studies documenting an inverse association between fetal growth and prenatal exposure to exogenous chemicals with renal clearance, such as PFOA.

We considered this hypothesis and its supporting scientific evidence in the context of our final conclusion from this review, and decided that it did not undermine our findings for two reasons. First, this hypothesis is not relevant to associations found in animal studies. In our review of PFOA, the experimental animal evidence was robust and mirrored the human evidence, lending support for the association between PFOA exposure and low birth weight (Koustas et al. 2014). Second, we systematically reviewed the literature for evidence of the relationship between birth weight and maternal glomerular filtration rate (see Supplemental Material, List of Studies Included in Systematic Review of the Relationship between Birth Weight and Maternal Glomerular Filtration Rate) and concluded that there is currently insufficient evidence to support the reverse causality hypothesis for associations between fetal growth and maternal glomerular filtration rate in humans. Additional research is needed to confirm or disprove this hypothesis. Thus, although we cannot disprove reverse causality, we have concluded that there is currently inconclusive evidence to justify altering our conclusions regarding the strength of human evidence. However, review authors were cognizant of the potential for these physiological factors associated with pregnancy to account for the negative association of PFOA with low birth weight. A preliminary study based on physiologically based pharmacokinetic (PBPK) modeling of a meta-analysis of seven published epidemiology studies suggested that a portion of the association between PFOA and low birth weight was attributed to confounding by GFR (Verner et al. 2014). Another study investigating hematologic changes and pregnancy outcomes similarly showed that low hemoglobin in late pregnancy was associated with low birth weight, but the association disappeared when adjusting for plasma volume (Whittaker et al. 1996). However,

there still remains a lack of human evidence that this is indeed the case for external chemical exposures. Although future studies may emerge with more conclusive evidence, we felt that although the reverse causation hypothesis is reasonable and warrants further investigation, without stronger evidence, and in light of the strength of the animal data, downgrading the final conclusion for 'sufficient' for the human evidence was not justifiable at this time.

Ultimately, our application of the Navigation Guide approach led to a clear and concise concluding statement, resulting from a systematic and transparent review of the literature developed from comprehensive and transparent methods used in the clinically sciences that have been demonstrated to reduce bias (Antman et al. 1996; Higgins and Green 2011). This is unique to the Navigation Guide systematic review methodology and the method under development by the National Toxicology Program (National Toxicology Program 2013; Rooney et al. 2014). A comparison of our results to those of previous reviews of PFOA (DeWitt et al. 2009; Hekster et al. 2003; Jensen and Leffers 2008; Kennedy et al. 2004; Kudo and Kawashima 2003; Lau et al. 2004; Lau et al. 2007; Lindstrom et al. 2011; Olsen et al. 2009; Post et al. 2012; Steenland et al. 2010; White et al. 2011) showed that the application of the Navigation Guide provided more transparency about the steps taken in the review and a consistent path to a clear answer compared to methods of expert-based narrative review that are currently employed in environmental health (Woodruff and Sutton 2014).

Adami et al. have proposed a framework to combine the toxicological and epidemiological evidence to establish causal inference (Adami et al. 2011; Simpkins et al. 2011). While similar to the Navigation Guide in seeking greater transparency overall in research synthesis and striving to integrate human and non-human evidence into a final conclusion, the methods differ in substantive, fundamental ways. Specifically, the Adami method does not conform to key features

of systematic review methodologies, i.e., an *a priori* protocol, a comprehensive search strategy, a risk of bias assessment, and data analysis. Moreover, whereas the Navigation Guide, as modeled after IARC (International Agency for Research on Cancer 2006) gives primacy to the strength of the human evidence stream in the absence of an established mode of action, in the Adami method conclusions about a body of evidence rests explicitly on whether or not a mode of action relevant to humans has been established by the toxicological evidence: i.e., if the mode of action established in animal models is considered to not be relevant to humans, then the biological plausibility of the effect observed in humans through the proposed mode of action is considered to be "highly unlikely." More research targeted on identifying and evaluating the utility, transparency, and robustness of different methods, including the questions they are suited for answering, will be useful in the future as the application of improved methods becomes more widespread (Krauth et al. 2013).

Limitations

One benefit of our adoption of the IARC approach is that it was transparent and simple to integrate the evidence from human and non-human bodies of available evidence once we rated each stream's strength of evidence separately. However, this meant that quantitative evaluations of the effect estimates for each body of evidence were kept separate and not integrated earlier on in the process. There has been much discussion recently in several research fields to utilize quantitative methods that can integrate diverse sources of data, such as human and non-human toxicity evidence, into a single quantitative model that can account for the different sources of data and expected contribution of each dataset to the evidence for human toxicity (DuMouchel and Harris 1983; Jones et al. 2009; Peters et al. 2005). Future investigation into methods to quantitatively integrate these diverse sources of data, for example in a hierarchical Bayesian

model, is warranted and would be an important contribution to advancing strength of evidence conclusions in environmental health.

The nomenclature of the overall strength of the human evidence, i.e., the terms 'known,' 'probably,' and 'toxic' generally had differing connotations among review authors despite agreement on the underlying definitions that supported the final conclusion. Some of the review authors found 'known to be toxic' to be an accurate descriptor of the body of evidence while others felt the descriptor 'probably toxic' was more appropriate. Our discussions of the variability of our own subjective reactions to 'known' and 'probably' emphasized the need for further delineation of *a priori* objective criteria for the strength of the evidence definitions.

Our different subjective reactions over terminology were resolved by focusing our discussion on the definitions we had established for each strength of evidence rating (Johnson et al. 2014; Koustas et al. 2014). From this discussion, ultimately all authors agreed with the final concluding statement. However, such consensus may not always be possible, as the available evidence is not always clear-cut. Conclusions about the strength of the evidence regarding toxicity must be made for regulatory purposes, for choosing less toxic alternatives, and/or for other purposes, and, as in the clinical sciences, complete agreement on the strength of the evidence should not be a criterion for enabling government agencies, professional societies, healthcare organizations, or others to make a determination. An example of this is Proposition 65 in the state of California, a voter-approved initiative that gives the State authority to classify chemicals deemed to cause cancer, birth defects, or reproductive health effects (California Office of Environmental Health Hazard Assessment 2013). One mechanism by which chemicals are added to the list is if either of two independent scientific committees concludes that the chemical has clearly been shown to cause these adverse health effects. Consensus is not required from both committees, and even

within an individual committee the vote to add a chemical to the list does not have to be unanimous—for example, the recent addition of Tris(1, 3-dichloro-2-propyl) phosphate (TDCPP) was determined based on a 5-1 vote in one committee (California Office of Environmental Health Hazard Assessment 2011).

Addressing a lack of consensus in the interpretation of scientific evidence reinforced a key methodological strength of systematic reviews, i.e., transparent definitions and documentation of the basis of a conclusion, so that the rationale for the final toxicity statement can be readily interpreted and/or contested by outside entities. In particular, it is critical to provide both a final recommendation and the documentation and justification leading to this conclusion. Additionally, we anticipate that readers seeing our concluding statement will have their own subjective connotations and reactions. While our nomenclature (i.e., known, possibly, etc.) was developed by modifying the nomenclature used by IARC and EPA for many years to classify carcinogens, the use in this context, adapted to be more broadly applicable to both carcinogens and non-carcinogens, and its utility to decision-makers are untested (International Agency for Research on Cancer 2006; U.S. Environmental Protection Agency 1991, 1996).

Specifically, there is currently no consensus in environmental health on how to name and communicate the strength of the evidence, and indeed there are many examples of similar terms that are commonly used to characterize varying strengths of evidence—for example, terms used to describe 'moderate' evidence include "balance of evidence," "balance of probabilities," "reasonable grounds of concern," and "strong possibility" (Gee 2008). Research related to climate change has shown that the public consistently misinterprets probabilistic statements such as *unlikely* or *very unlikely*, used in Intergovernmental Panel on Climate Change reports, and there are large individual differences in the interpretation of the statements which are associated

with the public's views and beliefs on climate change (Budescu et al. 2012). Research on better ways to communicate uncertainty is critical, and discussion of improved communication needs to include the users of the information, such as policy makers and the public.

This case study was limited to human and non-human animal data. There is a need to expand the scope of the Navigation Guide systematic review method to incorporate the results of *in vitro* studies and other modern methods of toxicology testing into the reviewed evidence stream. It is critical to develop such approaches as *in vitro* and other model systems and types of data will play an increasingly important role in the regulatory sphere as advances in technology allow for the rapid production of large quantities of data, such as those utilized in high-throughput screening (National Research Council 2007; U.S. Food and Drug Administration 2011).

Furthermore, our first case study of applying the Navigation Guide ended with Step 3, and we did not make a final recommendation about what to do about the science. The final Step 4 of the Navigation Guide is where the conclusion regarding toxicity is combined with additional information such as exposure prevalence, consideration of available alternatives, values and preferences to determine the final recommendation for public health protection. The Navigation Guide method allows for substances 'known to be toxic' to have discretionary recommendations, and substances 'possibly toxic' to have strong recommendations, depending on these and other potential factors. While we did not address this step for this case study due to resource limitations, carrying a case study through all the Navigation Guide steps is a research need for the future, as this will demonstrate how to apply the Navigation Guide in risk management decisions.

Lastly, exposures to environmental contaminants that lead to chronic disease or adverse reproductive and developmental health outcomes are complex and poorly understood. Such harm can be irreversible and can span across generations, making a strong case for timely decision-making and actions to prevent harm. However, having limited data or multiple studies of varying quality and findings can often hinder the ability to take such action. Developing criteria to evaluate diverse sources of scientific evidence in order to support action on the science is lacking, and therefore a critical unmet research need (Krauth et al. 2013).

Conclusion

Our case study demonstrates an application of the Navigation Guide to apply the rigor and transparency of systematic review methodology from the clinical sciences to make strength of evidence conclusions in environmental health. In this paper, we combined the strength of evidence ratings from the non-human (Koustas et al. 2014) and human evidence (Johnson et al. 2014) following the framework proposed in the Navigation Guide (Woodruff et al. 2011a) and review authors came to the final conclusion that "exposure to PFOA is 'known to be toxic' to human reproduction and development based on sufficient evidence of decreased fetal growth in both human and non-human mammalian species." This demonstrated the utility of the Navigation Guide to systematically evaluate the available evidence to answer questions relevant to environmental health. We anticipate that future applications of the Navigation Guide methodology to additional case studies will refine and improve the approach, contributing to the ultimate goal of supporting timely evidence-based decisions and recommendations for the prevention of harm to public health.

References

Adami HO, Berry SC, Breckenridge CB, Smith LL, Swenberg JA, Trichopoulos D, et al. 2011. Toxicology and epidemiology: Improving the science with a framework for combining toxicological and epidemiological evidence to establish causal inference. Toxicol Sci 122:223-234.

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. 1992. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 268:240-248.

Apelberg BJ, Goldman LR, Calafat AM, Herbstman JB, Kuklenyik Z, Heidler J, et al. 2007. Determinants of fetal exposure to polyfluoroalkyl compounds in Baltimore, Maryland. Environ Sci Technol 41:3891-3897.

Arbuckle TE, Kubwabo C, Walker M, Davis K, Lalonde K, Kosarac I, et al. 2012. Umbilical cord blood levels of perfluoroalkyl acids and polybrominated flame retardants. Int J Hyg Environ Health 216:184-194.

Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. 2011. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 64:401-406.

Bebarta V, Luyten D, Heard K. 2003. Emergency medicine animal research: Does use of randomization and blinding affect the results? Acad Emerg Med 10:684-687.

Bero L. 2013. Why the cochrane risk of bias tool should include funding source as a standard item [editorial]. Cochrane database of systematic reviews 12:ED000075.

Budescu DV, Por H-H, Broomell SB. 2012. Effective communication of uncertainty in the IPCC reports. Climatic change 113:181-200.

California Office of Environmental Health Hazard Assessment. 2011. A chemical listed effective October 28, 2011 as known to the state of California to cause cancer tris(1,3-dichloro-2-propyl) phosphate (TDCPP) (cas no. 13674-87-8). Available: http://oehha.ca.gov/prop65/prop65_list/102811list.html [accessed 20 February 2014].

California Office of Environmental Health Hazard Assessment. 2013. Safe drinking water and toxic enforcement act of 1986. Available: http://www.oehha.org/prop65/law/P65law72003.html [accessed 20 February 2014].

Deeks JJ, Higgins JPT, Altman DG (editors). 2011. Chapter 9:Analysing data and undertaking meta-analyses. In: Cochrane handbook for systematic reviews of interventions version 5.1.0, (Higgins JPT, Green S, eds) Available: http://www.cochrane-handbook.org [accessed 15 April 2013].

DeWitt JC, Shnyra A, Badr MZ, Loveless SE, Hoban D, Frame SR, et al. 2009. Immunotoxicity of perfluorooctanoic acid and perfluorooctane sulfonate and the role of peroxisome proliferator-activated receptor alpha. Crit Rev Toxicol 39:76-94.

DuMouchel WH, Harris JE. 1983. Bayes methods for combining the results of cancer studies in humans and other species. Journal of the American Statistical Association 78:293-308.

Fox DM. 2010. The convergence of science and governance: Research, health policy, and american states. Berkeley, CA:University of California Press.

Fromme H, Mosch C, Morovitz M, Alba-Alejandre I, Boehmer S, Kiranoglu M, et al. 2010. Pre- and postnatal exposure to perfluorinated compounds (PFCs). Environmental science & technology 44:7123-7129.

Gee D. 2008. Establishing evidence for early action: The prevention of reproductive and developmental harm. Basic & clinical pharmacology & toxicology 102:257-266.

GRADE Working Group. 2012. Available: http://www.gradeworkinggroup.org/ [accessed September 13 2012].

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. 2008. GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. BMJ 336:924-926.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011a. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol 64:1303-1310.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011b. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol 64:1294-1302.

Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. 2011c. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 64:1311–1316.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. 2011d. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol 64:1311–1316.

Halldorsson TI, Rytter D, Haug LS, Bech BH, Danielsen I, Becher G, et al. 2012. Prenatal exposure to perfluorooctanoate and risk of overweight at 20 years of age: A prospective cohort study. Environ Health Perspect 120:668.

Hekster FM, Laane RW, de Voogt P. 2003. Environmental and toxicity effects of perfluoroalkylated substances. Rev Environ Contam Toxicol 179:99-121.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. 2003. Measuring inconsistency in meta-analyses. BMJ 327:557-560.

Higgins JPT, Green S. 2011. Cochrane handbook for systematic reviews of interventions. Version 5.1.0 [updated march 2011].The Cochrane Collaboration. Available from http://www.cochrane-handbook.org

Institute of Medicine. 2007. Preterm birth: Causes, consequences, and prevention. Washington, DC:National Academy of Sciences.

Institute of Medicine, Eden J, Wheatley B, McNeil B, Sox H. 2008. Knowing what works in health care: A roadmap for the nation:National Academies Press.

International Agency for Research on Cancer. 2006. IARC monographs on the evaluation of carcinogenic risks to humans: Preamble (amended january 2006). Lyon, France.

Jensen AA, Leffers H. 2008. Emerging endocrine disrupters: Perfluoroalkylated substances. Int J Androl 31:161-169.

Johnson PI, Sutton P, Atchley DS, Koustas E, Lam J, Sen S, et al. 2014. The Navigation Guide - evidence-based medicine meets environmental health: Systematic review of human evidence for PFOA effects on fetal growth. Environ Health Perspect; http://dx.doi.org/10.1289/ehp.1307893 [Online 25 June 2014].

Jones DR, Peters JL, Rushton L, Sutton AJ, Abrams KR. 2009. Interspecies extrapolation in environmental exposure standard setting: A bayesian synthesis approach. Regul Toxicol Pharmacol 53:217-225.

Kennedy GL, Jr., Butenhoff JL, Olsen GW, O'Connor JC, Seacat AM, Perkins RG, et al. 2004. The toxicology of perfluorooctanoate. Crit Rev Toxicol 34:351-384.

Kim S, Choi K, Ji K, Seo J, Kho Y, Park J, et al. 2011. Trans-placental transfer of thirteen perfluorinated compounds and relations with fetal thyroid hormones. Environ Sci Technol 45:7465-7472.

Kim SK, Lee KT, Kang CS, Tao L, Kannan K, Kim KR, et al. 2011. Distribution of perfluorochemicals between sera and milk from the same mothers and implications for prenatal and postnatal exposures. Environmental Pollution 159:169-174.

Kimmel CA, Holson JF, Hogue CJ, Carlo G. 1984. Reliability of experimental studies for predicting hazards to human development. (NCTR Technical Report for Experiment No 6015). Jefferson, AR.

Koustas E, Lam J, Sutton P, Johnson PI, Atchley DS, Sen S, et al. 2014. The Navigation Guide - evidence-based medicine meets environmental health: Systematic review of nonhuman evidence for PFOA effects on fetal growth. Environ Health Perspect; http://dx.doi.org/10.1289/ehp.1307177 [Online 25 June 2014].

Krauth D, Woodruff TJ, Bero L. 2013. Instruments for assessing risk of bias and other methodological criteria of published animal studies: A systematic review. Environ Health Perspect 121:985-992.

Krauth D, Anglemyer A, Philipps R, Bero L. 2014. Nonindustry-sponsored preclinical study on statins yield greater efficacy estimates than industry-sponsored studies: A meta-analysis. PLOS Biol. 12(1):e1001770.

Kudo N, Kawashima Y. 2003. Toxicity and toxicokinetics of perfluorooctanoic acid in humans and animals. J Toxicol Sci 28:49-57.

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. 2012. A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490:187-191.

Lau C, Butenhoff JL, Rogers JM. 2004. The developmental toxicity of perfluoroalkyl acids and their derivatives. Toxicol Appl Pharmacol 198:231-241.

Lau C, Anitole K, Hodes C, Lai D, Pfahles-Hutchens A, Seed J. 2007. Perfluoroalkyl acids: A review of monitoring and toxicological findings. Toxicol Sci 99:366-394.

Lindstrom AB, Strynar MJ, Libelo EL. 2011. Polyfluorinated compounds: Past, present, and future. Environ Sci Technol 45:7954-7961.

Loccisano AE, Longnecker MP, Campbell Jr JL, Andersen ME, Clewell III HJ. 2013. Development of PBPK models for PFOA and PFOS for human pregnancy and lactation life stages. Journal of toxicology and environmental health, Part A 76:25-57.

Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. 2012. Industry sponsorship and research outcome. Cochrane Database of Systematic Reviews; doi: 10.1002/14651858.MR000033.pub2.

Macleod MR, O'Collins T, Howells DW, Donnan GA. 2004. Pooling of animal experimental data reveals influence of study design and publication bias. Stroke; a journal of cerebral circulation 35:1203-1208.

Mondal D, Lopez-Espinosa M-J, Armstrong B, Stein CR, Fletcher T. 2012. Relationships of perfluorooctanoate and perfluorooctane sulfonate serum concentrations between child-mother pairs in a population with perfluorooctanoate exposure from drinking water. Environ Health Perspect.

Monroy R, Morrison K, Teo K, Atkinson S, Kubwabo C, Stewart B, et al. 2008. Serum levels of perfluoroalkyl compounds in human maternal and umbilical cord blood samples. Environ Res 108:56-62.

National Research Council. 2007. Toxicity testing in the 21st century: A vision and strategy. Washington, DC:The National Academies Press.

National Research Council. 2009. Science and decisions: Advancing risk assessment. Washington, D.C.:National Academies Press.

National Research Council. 2011. Review of the environmental protection agency's draft IRIS assessment of formaldehyde. Washington, D.C.:National Academies Press.

National Research Council. 2014. Review of EPA's Integrated Risk Information System (IRIS) Process. Washington, D.C.:National Academies Press.

National Toxicology Program. 2013. OHAT implementation of systematic review. Available: http://ntp.niehs.nih.gov/?objectid=960B6F03-A712-90CB-8856221E90EDA46E [accessed November 7th 2013].

Nolan LA, Nolan JM, Shofer FS, Rodway NV, Emmett EA. 2009. The relationship between birth weight, gestational age and perfluorooctanoic acid (PFOA)-contaminated public drinking water. Reprod Toxicol 27:231-238.

Olsen GW, Butenhoff JL, Zobel LR. 2009. Perfluoroalkyl chemicals and human fetal development: An epidemiologic review with clinical and toxicological perspectives. Reprod Toxicol 27:212-230.

Peters JL, Rushton L, Sutton AJ, Jones DR, Abrams KR, Mugglestone MA. 2005. Bayesian methods for the cross-design synthesis of epidemiological and toxicological evidence. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54:159-172.

Post GB, Cohn PD, Cooper KR. 2012. Perfluorooctanoic acid (PFOA), an emerging drinking water contaminant: A critical review of recent literature. Environ Res 116:93-117.

Prevedouros K, Cousins IT, Buck RC, Korzeniowski SH. 2006. Sources, fate and transport of perfluorocarboxylates. Environmental Science & Technology 40:32-44.

Rennie D, Chalmers I. 2009. Assessing authority. JAMA 301:1819-1821.

Rooney AA, Boyles AL, Wolfe MS, Bucher JR, Thayer KA. 2014. Systematic review and evidence integration for literature-based environmental health science assessments. Environ Health Perspec DOI:10.1289/ehp.1307972

Savitz DA. 2007. Guest editorial: Biomarkers of perfluorinated chemicals and birth weight. Environ Health Perspect 115:A528-529.

Savitz DA, Stein CR, Bartell SM, Elston B, Gong J, Shin HM, et al. 2012. Perfluorooctanoic acid exposure and pregnancy outcome in a highly exposed community. Epidemiology 23:386-392.

Sawaya GF, Guirguis-Blake J, LeFevre M, Harris R, Petitti D. 2007. Update on methods: Estimating certainty and magnitude of net benefit. Ann Intern Med. 147: 871-875.

Simpkins JW, Swenberg JA, Weiss N, Brusick D, Eldridge JC, Stevens JT, et al. 2011. Atrazine and breast cancer: A framework assessment of the toxicological and epidemiological evidence. Toxicol Sci 123:441-459.

Steenland K, Fletcher T, Savitz DA. 2010. Epidemiologic evidence on the health effects of perfluorooctanoic acid (PFOA). Environ Health Perspect 118:1100-1108.

Stein CR, Savitz DA, Dougan M. 2009. Serum levels of perfluorooctanoic acid and perfluorooctane sulfonate and pregnancy outcome. Am J Epidemiol 170:837-846.

Sterne JAC, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. 2011. Recommendations for exmining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 343:d4002.

Sterne JAC. 2013. Why the cochrane risk of bias tool should not include funding source as a standard item [editorial]. Cochrane database of systematic reviews 12:ED000076.

Taubes G. 1995. Epidemiology faces its limits. Science 269:164-&.

U.S. Environmental Protection Agency. 1991. Guidelines for developmental toxicity risk assessment. Available: http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid=23162#Download [accessed 27 February 2014].

U.S. Environmental Protection Agency. 1996. Guidelines for reproductive toxicity risk assessment. Available: http://www.epa.gov/raf/publications/pdfs/REPRO51.PDF [accessed 27 February 2014].

U.S. Environmental Protection Agency. 2012. Basic information | PFOA and fluorinated telomers | oppt | oppts| US EPA. Available: http://epa.gov/oppt/pfoa/pubs/pfoainfo.html#background [accessed May 9 2013].

U.S. Food and Drug Administration. 2011. Advancing regulatory science at FDA: A strategic plan. Washington, D.C.:U.S. Department of Health and Human Services.

UCSF Program on Reproductive Health and the Environment. 2013. Application of the navigation guide: Case study of PFOA effects on fetal growth. Available: http://prhe.ucsf.edu/prhe/navigationguide.html [accessed 20 February 2014].

Verner M, Loccisano A, Yoon M, Wu H, McDougall R, Maisonet M., et al. 2014. The association between prenatal exposure to Perfluoroalkyl Substances (PFAS) and reduced birth weight: is golmerular filtration rate the underlying cause? Abstract 89 occurring in The Toxicologist, supplement to Toxicological Sciences. 138(1):18.

Viechtbauer W. 2010. Conducting meta-analyses in r with the metafor package. Journal of Statistical Software 36:1-48.

Viswanathan M, Ansari M, Berkman N, Chang S, Hartling L, McPheeters L, et al. 2012. Assessing the risk of bias of individual studies in systematic reviews of health care interventions. (Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews). AHRQ Publication No. 12-EHC047-EF.

White SS, Fenton SE, Hines EP. 2011. Endocrine disrupting properties of perfluorooctanoic acid. The Journal of Steroid Biochemistry and Molecular Biology 127:16-26.

Whittaker PG, Macphail S, Lind T. 1996. Serial hermatologic changes and pregnancy outcome. Obstetrics & Gynecology 88(1):33-39.

Whitworth KW, Haug LS, Baird DD, Becher G, Hoppin JA, Skjaerven R, et al. 2012. Perfluorinated compounds in relation to birth weight in the norwegian mother and child cohort study. Am J Epidemiol 175:1209-1216.

Woodruff TJ, Sutton P, The Navigation Guide Work Group. 2011a. An evidence-based medicine methodology to bridge the gap between clinical and environmental health sciences. Health Aff (Millwood) 30:931-937.

Woodruff TJ, Zota AR, Schwartz JM. 2011b. Environmental chemicals in pregnant women in the US: NHANES 2003-2004. Environ Health Perspect 119:878-885.

Woodruff TJ, Sutton P. 2014. The Navigation Guide: An improved method for translating environmental health science into better health outcomes. Environmenal Health Perspectives; http://dx.doi.org/10.1289/ehp.1307175 [Online 25 June 2014].

Table 1. Human and animal PECO statements.

PECO element	Human evidence	Animal evidence
Study question	Does developmental exposure to perflurooctanoic acid (PFOA) affect fetal growth in humans?	Does developmental exposure to perflurooctanoic acid (PFOA) affect fetal growth in animals?
Participants	Humans that are studied during reproductive/developmental time period (before and/or during pregnancy or development)	Animals from non-human species that are studied during reproductive/developmental time period (before and/or during pregnancy for females or during development for embryos)
Exposure	Exposure to perfluorooctanoic acid (PFOA), CAS# 335-67-1, or its salts during the time before pregnancy and/or during pregnancy for females or directly to fetuses	One or more oral, subcutaneous or other treatment(s) of any dosage with PFOA, CAS# 335-67-1, or its salts during the time before pregnancy and/or during pregnancy for females or directly to embryos
Comparators	Humans exposed to lower levels of PFOA than the more highly exposed humans	Experimental animals receiving different doses of PFOA or vehicle-only treatment
Outcomes	Effects on fetal growth, birth weight, and/or other measures of size, such as length	Changes in fetal weight near term (for example, embryonic day 18 for mice and embryonic day 21 for rat); birth weight; and/or other measures of size at term or birth, such as length.

Table 2. Summary table of the quality ratings given to each body of evidence.

Rating factor	**Human**	**Non-human mammalian**	**Non-mammalian**
Starting (initial) rating	Moderate	High	High
Downgrade factors			
Risk of bias across studies	0	-1	-1
Indirectness	0	0	-1
Inconsistency	0	0	0
Imprecision	0	0	0
Publication bias	0	0	0
Upgrade factors			
Large magnitude of effect	0	NA	NA
Dose response	0	NA	NA
Confounding minimizes effect	0	NA	NA
Overall grade	0	-1	-2
Resulting rating	Moderate	Moderate	Low

'0' – no change in rating; '-1' – decrease rating by 1 level; '-2' – decrease rating by 2 levels; NA – Not Applicable

Figure legends

Figure 1. Overview of process to rate the quality and strength of the evidence.

Figure 2. Flow chart of the progression from the literature search to inclusion in the systematic review and meta-analysis.

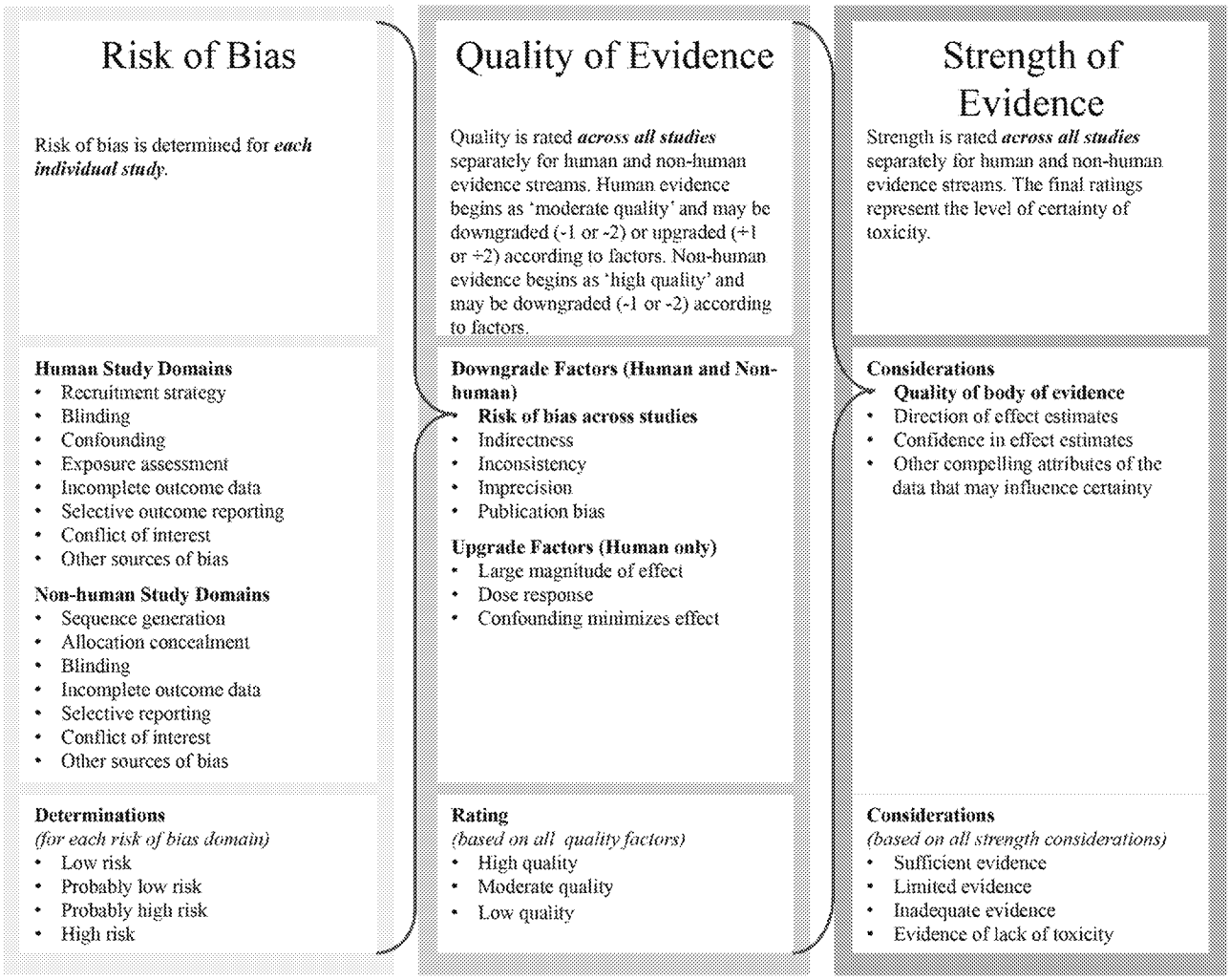

Figure 1

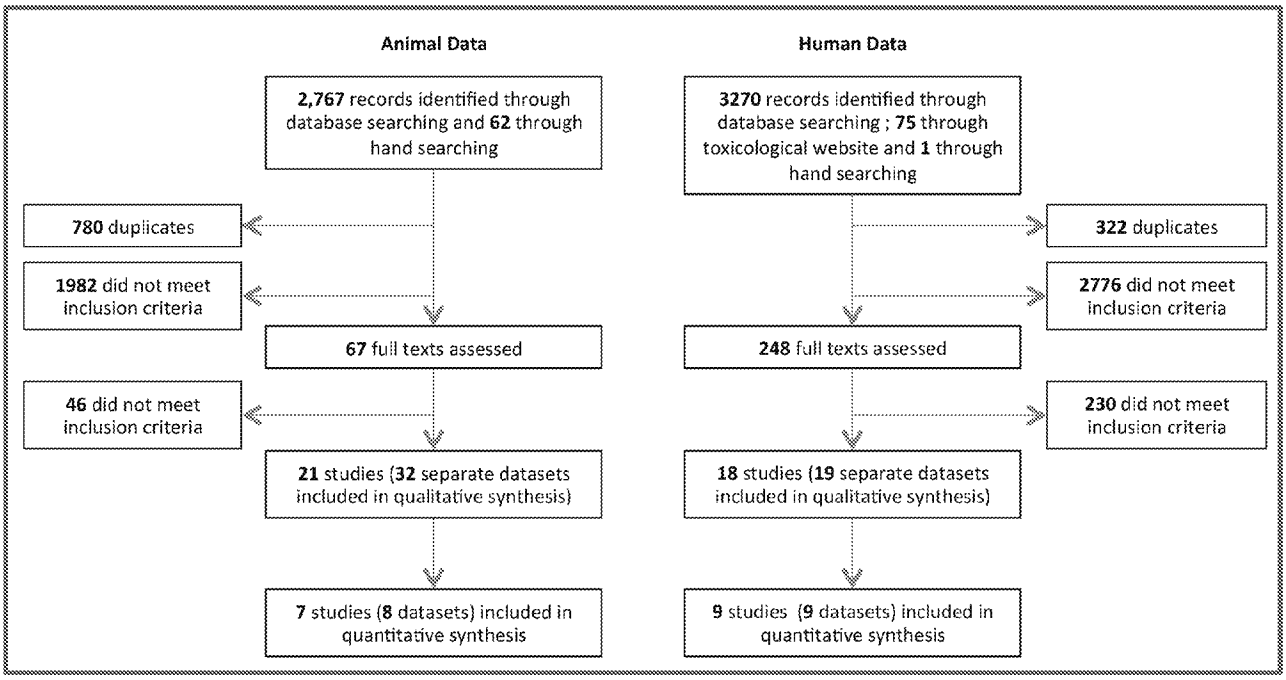
Animal Data
Human Data
2,767 records identified through database searching and 62 through hand searching
3270 records identified through database searching ; 75 through toxicological website and 1 through hand searching
780 duplicates
322 duplicates
1982 did not meet inclusion criteria
2776 did not meet inclusion criteria
67 full texts assessed
248 full texts assessed
46 did not meet inclusion criteria
230 did not meet inclusion criteria
21 studies (32 separate datasets included in qualitative synthesis)
18 studies (19 separate datasets included in qualitative synthesis)
7 studies (8 datasets) included in quantitative synthesis
9 studies (9 datasets) included in quantitative synthesis

Figure 2

Supplemental Material

The Navigation Guide—Evidence-Based Medicine Meets Environmental Health: Integration of Animal and Human Evidence for PFOA Effects on Fetal Growth

Juleen Lam, Erica Koustas, Patrice Sutton, Paula I. Johnson, Dylan S. Atchley, Saunak Sen, Karen A. Robinson, Daniel A. Axelrad, and Tracey J. Woodruff

Table of Contents **Page**

Navigation Guide Workgroup Members

Tracey J. Woodruff (University of California, San Francisco)

Patrice Sutton (University of California, San Francisco)

Vincent James Cogliano (International Agency for Research on Cancer)

Kate Guyton (Environmental Protection Agency [EPA])

Julia Quint (Retired, California Department of Public Health)

Lauren Zeise (CA-EPA)

Judith Balk (University of Pittsburgh [UP])

Lisa Bero (UCSF)

Jeanne Conry (Kaiser Permanente, American College of Obstetricians and Gynecologists District IX)

Daniel M. Fox (Emeritus, Milbank Memorial Fund)

David Gee (European Environmental Agency)

Rivka Gordon (Association of Reproductive Health Professionals [ARHP])

Sarah Janssen (Natural Resources Defense Council)

Beth Jordan (ARHP)

Victoria Maizes (University of Arizona)

Mark Miller (UCSF)

Michele Ondeck (UP)

Karen Pierce (San Francisco Department of Public Health)

Pablo Rodriguez (Brown Medical School and Women & Infants Hospital of Rhode Island)

Heather Sarantis (Collaborative on Health and the Environment)

Ted Schettler (Science and Environmental Health Network)

Sandy Worthington (Planned Parenthood Federation of America)

Instructions for rating the quality and strength of human and non-human evidence

1. Co-authors will independently review the final data and independently rate the quality of evidence according to *a priori* criteria set forth in the protocol.

2. Co-authors will compare their results. Any discrepancies between the co-authors' decisions will be resolved through discussion. The senior author (TW) will be the ultimate arbiter of the discrepancies that cannot be resolved through consensus among the co-authors. The final judgments of all reviewers will be documented.

3. The initial quality level of non-human experimental data is considered "high" consistent with GRADE guidelines for rating experimental studies (i.e., randomized controlled trials). The initial quality level of human observational data is considered "moderate". This is in contrast to GRADE guidelines, developed for clinical interventions, which assign observational studies an initial rating of "low" quality (Balshem et al. 2011). There is variability in the quality of studies, however, and not all observational studies may be low quality (Viswanathan et al. 2012). In environmental health, human observational data are the "best" data available for decision-making, and in this regard they are comparable to human randomized controlled trials (RCTs) in the clinical sciences. Because ethics virtually precludes human RCTs in environmental health, beginning human observational studies at "moderate" quality captures the value of these data relative to what data are available. In addition, human observational studies are recognized as being a reliable source of evidence in the clinical sphere, as not all healthcare decisions are, or can be, based on RCTs; (Institute of Medicine et al. 2008) recognition of the absolute value of human observational data evidence-based to clinical decision-making is also increasing (Peterson 2008; Halvorson 2008).

4. "Fetal growth" is the outcome being assessed in this review.

    - In humans, the outcome fetal growth includes all the following measures: birth weight, birth length, head circumference, and ponderal index; all of these measures are sufficiently similar to rate together as a measure of the same outcome.

    - In non-human mammalians, the outcome "fetal" growth includes all the following measures: "Fetal" data, which refers to when outcome measurements are taken from progeny near-term (i.e., E18 for mice, E21 for rats). "Pup" data, which refers to when outcome measurements are taken from progeny at or soon after birth.

    - In non-human non-mammalians, the outcome fetal growth is equivalent to "embryonic" growth and includes measures of weight, length or volume, depending on the model system.

5. For the purpose of the PFOA case study, there are 3 populations for which we are rating the quality of evidence for PFOA's effect on fetal growth: (1) the quality of human evidence for fetal growth; (2) the quality of mammalian animal evidence for fetal growth; and (3) the quality of non-human, non-mammalian evidence for fetal growth.

6. There are 5 categories that can lead to **downgrading** quality of evidence for an outcome: risk of bias (study limitations); indirectness; inconsistency; imprecision; and publication bias. According to GRADE, these 5 categories address nearly all issues that bear on the quality of evidence (Balshem et al. 2011). GRADE states that these categories were arrived at through a case- based process by members of GRADE, who identified a broad range of issues and factors related to the assessment of the quality of studies. All potential factors were considered, and through an iterative process of discussion and review, concerns were scrutinized and solutions narrowed by consensus to these five categories. GRADE also defines 3 categories that can lead to upgrading quality of evidence for an outcome: large effect; confounding would minimize effect; and dose response.

7. While GRADE specifies systematic review authors consider quality of evidence under a number of discrete categories and to either rate down or not on the basis of each category, they also state that rigid adherence to this approach ignores the fact that quality is actually a continuum and that an accumulation of limitations across categories can ultimately provide the impetus for rating down in quality (Guyatt et al. 2011a). Thus authors who decide to rate down quality by a single level will specify the one category most responsible for their decision while documenting all factors that contributed to the final decision to rate down quality.

8. The quality of evidence rating for human and non-human data will be translated into strength of evidence ratings for each stream of evidence.

9. The strength of evidence for human and non-human data will be combined into an overall statement of toxicity, i.e., known to be toxic to fetal growth; probably toxic to fetal growth; possibly toxic to fetal growth; known to be not-toxic to fetal growth.

I. Rate the Quality of Evidence

Each of the categories to consider in downgrading or upgrading the evidence is described in detail, below. Please record your results on the chart at the end of each category, including a brief explanation for your ratings.

Category 1. Rate the Quality of Study Limitations (Risk of Bias) (Guyatt et al. 2011b)

Possible ratings: 0=no change; -1 or -2 downgrade 1 or 2 levels.

The evidence from studies can be rated down if most of the relevant evidence comes from studies that suffer from a high risk of bias. Risk of bias is rated by outcome across studies. Study limitations for each outcome for individual studies and across studies are summarized in the heat maps.

GRADE outlines the following principles for moving from risk of bias in individual studies to rating quality of evidence across studies.

1. In deciding on the overall quality of evidence, one does not average across studies (for instance if some studies have no serious limitations, some serious limitations, and some very serious limitations, one does not automatically rate quality down by one level because of an average rating of serious limitations). Rather, judicious consideration of the contribution of each study, with a general guide to focus on the high-quality studies is warranted.

   (Note: Limitations to GRADE's risk of bias assessments as stated by GRADE: "First, empirical evidence supporting the criteria is limited. Attempts to show systematic difference between studies that meet and do not meet specific criteria have shown inconsistent results. Second, the relative weight one should put on the criteria remains uncertain. The GRADE approach is less comprehensive than many systems, emphasizing simplicity and parsimony over completeness. GRADE's approach does not provide a quantitative rating of risk of bias. Although such a rating has advantages, we share with the Cochrane Collaboration methodologists a reluctance to provide a risk of bias score that, by its nature, must make questionable assumptions about the relative extent of bias associated with individual items and fails to consider the context of the individual items.")

2. This judicious consideration requires evaluating the extent to which each study contributes toward the estimate of magnitude of effect. This contribution will usually reflect study sample size and number of outcome events larger studies with many events will contribute more, much larger studies with many more events will contribute much more.

3. One should be conservative in the judgment of rating down. That is, one should be confident that there is substantial risk of bias across most of the body of available evidence before one rates down for risk of bias.

4. The risk of bias should be considered in the context of other limitations. If, for instance, reviewers find themselves in a close-call situation with respect to two quality issues (risk of bias and, say, precision), GRADE suggests rating down for at least one of the two.

5. Notwithstanding the first four principles, reviewers will face close-call situations. You should acknowledge that you are in such a situation, make it explicit why you think this is the case, and make the reasons for your ultimate judgment apparent.

Type of study	Risk of bias (study limitations) rating	Rationale for your judgment
Human		
Non-human mammalian		
Non-human non-mammalian		

Category 2. Rate Indirectness of Evidence

Possible ratings: 0=no change; -1 or -2 downgrade 1 or 2 levels.

Quality of evidence (your confidence in estimates of effect) may decrease when substantial differences exist between the population, the exposure, or the outcomes measured in research studies under consideration in the review.

Evidence is direct when it directly compares the exposures in which we are interested when applied to the populations in which we are interested and measures outcomes important to the study question (in GRADE the outcomes must be important to patients).

Based on GRADE (Guyatt et al. 2011c) (as modified to reflect our "PECO" instead of "PICO" question), evidence can be indirect in one of three ways. (Note: GRADE includes a fourth type of indirectness that occurs when there are no direct (i.e., head-to-head) comparisons between two or more interventions of interest. This criterion is not relevant to our study question related to toxicity of PFOA; it could be relevant to future case studies.)

1. The population studied differs from the population of interest (the term applicability is often used for this form of indirectness). **Please note the Navigation Guide's *a priori* assumption is that mammalian evidence of a health effect/lack of health effect is deemed to be direct evidence of human health with regards to directness of the population**. This is a marked departure from GRADE (note: According to GRADE, in general, GRADE rates animal evidence down two levels for indirectness. They note that animal studies may, however, provide an important indication of drug toxicity. GRADE states, "Although toxicity data from animals does not reliably predict toxicity in humans, evidence of animal toxicity should engender caution in recommendations." However, GRADE does not preclude rating non-human evidence as high quality. They state, "Another type of nonhuman study may generate high- quality evidence. Consider laboratory evidence of change in resistance patterns of bacteria to antimicrobial agents (e.g., the emergence of methicillin-resistant staphylococcus aureus-MRSA). These laboratory findings may constitute high-quality evidence for the superiority of antibiotics to which MRSA is sensitive vs. methicillin as the initial treatment of suspected staphylococcus sepsis in settings in which MRSA is highly prevalent"), based on empirical evidence in environmental health science that the reliability of experimental animal (mammalian) data for reproductive and developmental health has been well established though multiple studies of concordance between mammalian animals and humans after exposure to a variety of chemical agents (Hemminki and Vineis 1985; Nisbet and Karch 1983; Kimmel et al. 1984; Nemec et al. 2006; Newman et al. 1993). Presently, there is no example of a chemical agent that has adversely affected human reproduction or development

but has not caused the same or similar adverse effects in mammalian animal models (Kimmel et al. 1984). The National Academy of Sciences (NAS) has recognized the importance of animal data in identifying potential developmental risks. According to the NAS, studies of comparison between developmental effects in animals and humans find that "there is concordance of developmental effects between animals and humans and that humans are as sensitive or more sensitive than the most sensitive animal species (National Research Council (U.S.) Committee on Developmental Toxicology and National Research Council (U.S.) Commission on Life Sciences 2000)." GRADE states that in general, one should not rate down for population differences unless one has compelling reason to think that the biology in the population of interest is so different than the population tested that the magnitude of effect will differ substantially. According to GRADE, most often, this will not be the case. In applying this GRADE principle to the Navigation Guide, non-human evidence would be rated down as indirect when it is a biologically inappropriate non-human model system for the health outcome under study.

2. The intervention (exposure) tested may differ from the exposure of interest, i.e., a difference in the chemical, route and/or dose. Decisions regarding indirectness of populations and exposure depend on an understanding of whether biological or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect. GRADE also states, "As with all other aspects of rating quality of evidence, there is a continuum of similarity of the intervention that will require judgment. It is rare, and usually unnecessary, for the intended populations and interventions to be identical to those in the studies, and we should only rate down if the differences are considered sufficient to make a difference in outcome likely."

3. Outcomes may differ from those of primary interest, for instance, surrogate outcomes that are not themselves important, but measured in the presumption that changes in the surrogate reflect changes in an important outcome. The difference between desired and measured outcomes may relate to time frame. When there is a discrepancy between the time frame of measurement and that of interest, whether to rate down by one or two levels will depend on the magnitude of the discrepancy. Another source of indirectness related to measurement of outcomes is the use of substitute or surrogate endpoints in place of the exposed population's important outcome of interest. In general, the use of a surrogate outcome requires rating down the quality of evidence by one, or even two, levels. Consideration of the biology, mechanism, and natural history of the disease can be helpful in making a decision about indirectness. Surrogates that are closer in the putative causal pathway to the adverse outcomes warrant rating down by only one level for indirectness. GRADE states that rarely, surrogates are sufficiently well established that one should choose not to rate down quality of evidence for indirectness. In general, evidence based on surrogate outcomes should usually trigger rating down, whereas the other types of indirectness will require a more considered judgment.

Type of study	**Indirectness rating**	**Rationale for your judgment**
Human		
Non-human mammalian		
Non-human non-mammalian		

Category 3. Rate Inconsistency of Evidence

Possible ratings: 0=no change; -1 or -2 downgrade 1 or 2 levels.

According to Cochrane, "when studies yield widely differing estimates of effect (heterogeneity or variability in results) investigators should look for robust explanations for that heterogeneity. … When heterogeneity exists and effects the interpretation of results, but authors fail to identify a plausible explanation, the quality of the evidence decreases."

Based on GRADE (Guyatt et al. 2011d), **a body of evidence is not rated up in quality if studies yield consistent results**, **but may be rated down in quality if inconsistent**. Their stated reason is that a consistent bias will lead to consistent, spurious findings.

GRADE suggests rating down the quality of evidence if large inconsistency (heterogeneity) in study results remains after exploration of a priori hypotheses that might explain heterogeneity. Judgment of the extent of heterogeneity is based on similarity of point estimates, extent of overlap of confidence intervals, and statistical criteria. GRADE's recommendations refer to inconsistencies in effect size, specifically to relative measures (risk ratios and hazard ratios or odds ratios), not absolute measures.

Based on GRADE, reviewers should consider rating down for inconsistency when:

1. Point estimates vary widely across studies;
2. Confidence intervals (CIs) show minimal or no overlap;
3. The statistical test for heterogeneity-which tests the null hypothesis that all studies in a meta-analysis have the same underlying magnitude of effect- shows a low P-value;
4. The I^2 -which quantifies the proportion of the variation in point estimates due to among-study differences-is large. (I.e., the I^2 index quantifies the degree of heterogeneity in a meta-analysis).

GRADE states that inconsistency is important **only when it reduces confidence in results in relation to a particular decision**. Even when inconsistency is large, it may not reduce confidence in results regarding a particular decision. For example, studies that are inconsistent related to the magnitude of a beneficial or harmful effect (but are in the same direction) would not be rated down; in instances when results are inconsistent as to whether there is a benefit or harm of treatment, GRADE would rate down the quality of evidence as a result of variability in results, because the meaning of the inconsistency is so relevant to the decision to treat or not to treat.

Type of study	Inconsistency rating	Rationale for your judgment
Human		
Non-human mammalian		
Non-human non-mammalian		

Category 4. Rate Imprecision of Evidence

Possible ratings: 0=no change; -1 or -2 downgrade 1 or 2 levels.

Cochrane states that when studies have few participants and few events, and thus have wide confidence intervals (CIs), authors can lower their rating of the quality of evidence. These ratings of precision are made as judgments by review authors.

GRADE defines evidence quality differently for systematic reviews and guidelines. For systematic reviews, quality refers to confidence in the estimates of effect. For guidelines, quality refers to the extent to which confidence in the effect estimate is adequate to support a particular decision (Guyatt et al. 2011). For the purpose of step 3 of Navigation Guide, we will use the systematic review definition, because the decision phase does not occur until step 4 when recommendations for prevention are made. Thus, when reviewing the data for imprecision, evaluate your confidence in the estimate of the effect.

According to GRADE, to a large extent, CIs inform the impact of random error on evidence quality. When considering the quality of evidence, the issue is whether the CI around the estimate of exposure effect is sufficiently narrow. If it is not, GRADE rates down the evidence quality by one level (for instance, from high to moderate). If the CI is very wide, GRADE might rate down by two levels.

Type of study	Imprecision rating	Rationale for your judgment
Human		
Non-human mammalian		
Non-human non-mammalian		

Category 5. Rate Publication Bias

Possible ratings: 0=no change; -1 or -2 downgrade 1 or 2 levels.

GRADE (Guyatt et al. 2011b) and Cochrane (Higgins and Green 2011) assess publication bias in a similar manner. Whereas "selective outcome reporting" is assessed for each study included in the review as part of the risk of bias assessment, "publication bias" is assessed on the body of evidence. GRADE states that "when an entire study remains unreported and the results relate to

the size of the effect- publication bias- one can assess the likelihood of publication bias only by looking at a group of studies."

Cochrane's definition of publication bias is "the *publication* or *non-publication* of research findings depending on the nature and direction of the results." Cochrane and GRADE are primarily concerned with *overestimates* of true effects of treatments or pharmaceuticals, especially related to "small studies effects", i.e., the tendency for estimates of an intervention to be more beneficial in smaller studies. There is empirical evidence in the clinical sciences that publication and other reporting biases result in over estimating the effects of interventions (Higgins and Green 2011).

In contrast, with the Navigation Guide, we are primarily concerned with *underestimating* the true effects of a chemical exposure, since in many cases population wide exposure has already occurred. Applying this inverted concern to GRADE's assessment for publication bias, leads to these considerations when rating publication bias:

- Early *negative* studies, particularly if small in size, are suspect. (GRADE is concerned with early *positive* studies).
- Authors of systematic reviews should suspect publication bias when studies are uniformly small, particularly when sponsored by the industry. (Same as GRADE)
- Empirical examination of patterns of results (e.g., funnel plots) may suggest publication bias but should be interpreted with caution. (Same as GRADE)
- More compelling than any of these theoretical exercises is authors' success in obtaining the results of some unpublished studies and demonstrating that the published and unpublished data show different results. (Same as GRADE)
- Comprehensive searches of the literature including unpublished studies, i.e., the grey literature, and a search for research in other languages are important to addressing publication bias. Note that Cochrane also states "comprehensive searching is not sufficient to prevent some substantial potential biases.

Type of study	**Publication bias rating**	**Rationale for your judgment**
Human		
Non-human mammalian		
Non-human non-mammalian		

Category 6. Rate Factors that Can Increase Quality of Evidence

Possible ratings: 0=no change; +1 or +2 upgrade 1 or 2 levels.

GRADE states that the circumstances for upgrading likely occur infrequently and are primarily relevant to observational and other non-randomized studies. Although it is possible to rate up results from randomized controlled trials, GRADE has yet to find a compelling circumstance for doing so (Guyatt et al. 2011e).

GRADE specifies 3 categories for increasing the quality of evidence (Guyatt et al. 2011e):

1. Large magnitude of effect. Modeling studies suggests that confounding (from non-random allocation) alone is unlikely to explain associations with a relative risk (RR) greater than 2 (or less than 0.5), and very unlikely to explain associations with an RR greater than 5 (or less than 0.2). Thus, these are the definitions of "large magnitude of effect" used to upgrade 1 or 2 levels, respectively. Also, GRADE is more likely to rate up if the effect is rapid and out of keeping with prior trajectory; usually supported by indirect evidence. GRADE presents empirical evidence to support these conclusions, and states that "although further research is warranted, both modeling and empirical work suggest the size of bias from confounding is unpredictable in direction but bounded in size. Hence, the GRADE group has previously suggested guidelines for rating quality of evidence up by one category (typically from low to moderate) for associations greater than 2, and up by two categories for associations greater than 5."

2. Dose-response gradient. Possible considerations include consistent dose response gradients in one or multiple studies, and/or dose response across studies, depending on the overall relevance to the body of evidence.

3. All plausible residual confounders or biases would reduce a demonstrated effect, or suggest a spurious effect when results show no effect. GRADE provides the following example of grading up evidence when observational studies have failed to demonstrate an association. Observational studies failed to confirm an association between vaccination and autism. This lack of association occurred despite the empirically confirmed bias that parents of autistic children diagnosed after the publicity associated with the article that originally suggested this relationship would be more likely to remember their vaccine experience than parents of children diagnosed before the publicity and presumably, than parents of non-autistic children. The negative findings despite this form of recall bias suggest rating up the quality of evidence.

Type of study	**Large magnitude of effect rating**	**Rationale for your judgment**
Human		

The results of the reviewers' ratings by population will be compiled and discussed leading to a final decision on overall quality of human evidence. The rationale for the decision will be fully documented.

1. Final decision on overall quality of human evidence:

(Example: Moderate quality is upgraded 1 step to high for Xyz reason(s))

---- High

---- Moderate

---- Low

---- Very

2. Final decision on overall quality of non-human mammalian evidence:

(Example: High quality is downgraded 1 step to moderate for Xyz reason(s))

---- High

---- Moderate

---- Low

---- Very

3. Final decision on overall quality of non-human non-mammalian evidence:

(Example: High quality is downgraded 1 step to moderate for Xyz reason(s))

---- High

---- Moderate

---- Low

---- Very

II. Rate the Strength of Evidence

The evidence quality ratings will be translated into strength of evidence for each population based on a combination of four criteria: (1) Quality of body of evidence; (2) Direction of effect; (3) Confidence in effect; and (4) Other compelling attributes of the data that may influence certainty (Figures 2 and 3). These strength of evidence ratings are linked to Tables 1 and 2, below, where their meaning is defined.

III. Combine Strength of Evidence For Human and Non-human Evidence

The final step in the process is to combine the strength of the evidence according to the chart in Figure 1. Combining the strength of evidence for human and non-human data will be produce an

overall statement of toxicity, i.e., known to be toxic to fetal growth; probably toxic to fetal growth; possibly toxic to fetal growth; known to be not-toxic to fetal growth.

List of studies included in systematic review of the relationship between birth weight and maternal glomerular filtration rate

Observational human studies—fetal growth and glomerular filtration rate

1) Akahori Y, Masuyama H, Hiramatsu Y. 2012. The correlation of maternal uric acid concentration with small-for-gestational-age fetuses in normotensive pregnant women. Gynecol Obstet Invest 73(2): 162-167.
2) Davison JM, Hytten FE. 1974. Glomerular filtration during and after pregnancy. J Obstet Gynaecol Br Commonw 81(8): 588-595.
3) Dunlop W, Furness C, Hill LM. 1978. Maternal haemoglobin concentration, haematocrit and renal handling of urate in pregnancies ending in the births of small-for-dates infants. Br J Obstet Gynaecol 85(12): 938-940.
4) Duvekot JJ, Cheriex EC, Pieters FA, Menheere PP, Schouten HJ, Peeters LL. 1995. Maternal volume homeostasis in early pregnancy in relation to fetal growth restriction. Obstet Gynecol 85(3): 361-367.
5) Faupel-Badger JM, Hsieh CC, Troisi R, Lagiou P, Potischman N. 2007. Plasma volume expansion in pregnancy: implications for biomarkers in population studies. Cancer Epidemiol Biomarkers Prev 16(9): 1720-1723.
6) Gibson HM. 1973. Plasma volume and glomerular filtration rate in pregnancy and their relation to differences in fetal growth. J Obstet Gynaecol Br Commonw 80(12): 1067-1074.
7) Knopp RH, Bergelin RO, Wahl PW, Walden CE. 1985. Relationships of infant birth size to maternal lipoproteins, apoproteins, fuels, hormones, clinical chemistries, and body weight at 36 weeks gestation. Diabetes 34 Suppl 2: 71-77.
8) Laughon SK, Catov J, Roberts JM. 2009. Uric acid concentrations are associated with insulin resistance and birthweight in normotensive pregnant women. Am J Obstet Gynecol 201(6): 582 e581-586.

Observational human studies—fetal growth and plasma volume expansion

9) Bernstein IM, Wulfkuhle K, Schonberg A. 2010. Fetal growth restriction is associated with reduced maternal plasma volume expansion. Abstract. Reproductive Sciences 1): 103A.
10) Blankson ML, Goldenberg RL, Cutter G, Cliver SP. 1993. The Relationship between Maternal Hematocrit and Pregnancy Outcome - Black-White Differences. Journal of the National Medical Association 85(2): 130-134.
11) Boomer AL, Christensen BL. 1982. Antepartum hematocrit, maternal smoking and birth weight. J Reprod Med 27(7): 385-388.

12) Dunlop W, Furness C, Hill LM. 1978. Maternal haemoglobin concentration, haematocrit and renal handling of urate in pregnancies ending in the births of small-for-dates infants. Br J Obstet Gynaecol 85(12): 938-940.
13) Gallery EDM, Saunders DM, Hunyor SN, Gyory AZ. 1979. Relationship between plasma-volume expansion and intra-uterine fetal growth in normal and hypertensive pregnancy. Abstract. Aust N Z J Obstet Gynaecol 19(3): 179-179.
14) Gibson HM. 1973. Plasma volume and glomerular filtration rate in pregnancy and their relation to differences in fetal growth. J Obstet Gynaecol Br Commonw 80(12): 1067-1074.
15) Hays PM, Cruikshank DP, Dunn LJ. 1985. Plasma volume determination in normal and preeclamptic pregnancies. Am J Obstet Gynecol 151(7): 958-966.
16) Hutchins CJ. 1980. Plasma volume changes in pregnancy in Indian and European primigravidae. Br J Obstet Gynaecol 87(7): 586-589.
17) Hytten FE, Stewart AM, Palmer JH. 1963. The relation of maternal heart size, blood volume and stature to the birth weight of the baby. J Obstet Gynaecol Br Commonw 70: 817-820.
18) Hytten FE, Paintin DB. 1963. Increase in plasma volume during normal pregnancy. J Obstet Gynaecol Br Emp 70: 402-407.
19) Keet MP, Jaroszewicz AM, van Schalkwyk DJ, Deale CJ, Odendaal HJ, Malan C, et al. 1981. Small-for-age babies: etiological factors in the Cape colored population. S Afr Med J 60(5): 199-203.
20) Lu ZM, Goldenberg RL, Cliver SP, Cutter G, Blankson M. 1991. The Relationship between Maternal Hematocrit and Pregnancy Outcome. Obstetrics and Gynecology 77(2): 190-194.
21) Pirani BB, Campbell DM, MacGillivray I. 1973. Plasma volume in normal first pregnancy. J Obstet Gynaecol Br Commonw 80(10): 884-887.
22) Rajalakshmi K, Raman L. 1985. Plasma volume changes in Indian women with normal pregnancy. Indian J Med Res 82: 521-527.
23) Rosso P, Donoso E, Braun S, Espinoza R, Fernandez C, Salas SP. 1993. Maternal hemodynamic adjustments in idiopathic fetal growth retardation. Gynecol Obstet Invest 35(3): 162-165.
24) Sagen N, Nilsen ST, Kim HC, Bergsjo P, Koller O. 1984. Maternal hemo globin concentration is closely related to birth weight in normal pregnancies. Acta Obstetricia et Gynecologica Scandinavica 63(3): 245-248.
25) Salas SP, Rosso P, Espinoza R, Robert JA, Valdes G, Donoso E. 1993. Maternal plasma-volume expansion and hormonal changes in women with idiopathic fetal growth-retardation. Obstetrics and Gynecology 81(6): 1029-1033.
26) Salas SP, Rosso P. 1998. Plasma volume, renal function, and hormonal levels in pregnant women with idiopathic fetal growth restriction or preeclampsia. Hypertension in Pregnancy 17(1): 69-79.
27) Scanlon KS, Yip R, Schieve LA, Cogswell ME. 2000. High and low hemoglobin levels during pregnancy: differential risks for preterm birth and small for gestational age. Obstet Gynecol 96(5 Pt 1): 741-748.

28) Steer P, Alam MA, Wadsworth J, Welch A. 1995. Relation between Maternal Hemoglobin Concentration and Birth-Weight in Different Ethnic-Groups. Br Med J 310(6978): 489-491.
29) Ueland K. 1976. Maternal cardiovascular dynamics. VII. Intrapartum blood volume changes. Am J Obstet Gynecol 126(6): 671-677.
30) von Tempelhoff GF, Heilmann L, Rudig L, Pollow K, Hommel G, Koscielny J. 2008. Mean maternal second-trimester hemoglobin concentration and outcome of pregnancy: a population-based study. Clin Appl Thromb Hemost 14(1): 19-28.
31) Whittaker PG, Macphail S, Lind T. 1996. Serial hematologic changes and pregnancy outcome. Obstet Gynecol 88(1): 33-39.
32) Gibson HM. 1973. Plasma volume and glomerular filtration rate in pregnancy and their relation to differences in fetal growth. J Obstet Gynaecol Br Commonw 80(12): 1067-1074.
33) Varga I, Rigo J, Jr., Somos P, Joo JG, Nagy B. 2000. Analysis of maternal circulation and renal function in physiologic pregnancies; parallel examinations of the changes in the cardiac output and the glomerular filtration rate. J Matern Fetal Med 9(2): 97-104.

Observational non-human mammalian studies—fetal growth and glomerular filtration rate

34) Abeni F, Bergoglio G, Masoero G, Terzano GM, Allegrini S. 2004. Plasma hormones and metabolites in Piedmontese cows during late pregnancy: relationships with calf birth weight. J Anim Sci 82(2): 438-444.

Observational non-human mammalian studies—fetal growth and plasma volume expansion

35) Haynes DM. 1954. Uterine blood volume. Obstet and Gynecol 3((5)): 517-522.

Experimental non-human mammalian studies—fetal growth and plasma volume expansion

36) Rumball CW, Bloomfield FH, Harding JE. 2008. Cardiovascular adaptations to pregnancy in sheep and effects of periconceptional undernutrition. Placenta 29(1): 89-94.
37) Van Mieghem T, van Bree R, Van Herck E, Deprest J, Verhaeghe J. 2009. Insulin-like growth factor-II regulates maternal hemodynamic adaptation to pregnancy in rats. Am J Physiol Regul Integr Comp Physiol 297(5): R1615-1621.

References

Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. 2011. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 64(4): 401-406.

Guyatt G, Oxman A, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. 2011. GRADE guidelines: 6. Rating the quality of evidence--imprecision. J Clin Epidemiol 64(12): 1283-1293.

Guyatt GH, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. 2011a. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 64(4): 383-394.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. 2011b. GRADE guidelines: 5. Rating the quality of evidence-publication bias. J Clin Epidemiol.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011c. GRADE guidelines: 8. Rating the quality of evidence-indirectness. J Clin Epidemiol.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. 2011d. GRADE guidelines: 7. Rating the quality of evidence-inconsistency. J Clin Epidemiol.

Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. 2011e. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol.

Halvorson GC. 2008. Electronic medical records and the prospect of real time evidence development. In: Evidence-based medicine and the changing nature of health care: 2007 IOM annual meeting summary. Washington, DC:The National Academies Press.

Hemminki K, Vineis P. 1985. Extrapolation of the evidence on teratogenicity of chemicals between humans and experimental animals: chemicals other than drugs. Teratog Carcinog Mutagen 5(4): 251-318.

Higgins JPT, Green S. 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 The Cochrane Collaboration.

Institute of Medicine, Eden J, Wheatley B, McNeil B, Sox H. 2008. Knowing what works in health care: a roadmap for the nation: National Academies Press.

Kimmel CA, Holson JF, Hogue CJ, Carlo G. 1984. Reliability of Experimental Studies for Predicting Hazards to Human Development. (NCTR Technical Report for Experiment No 6015). Jefferson, AR.

National Research Council (U.S.) Committee on Developmental Toxicology, National Research Council (U.S.) Commission on Life Sciences. 2000. Scientific frontiers in developmental toxicology and risk assessment. Washington, DC: National Academy Press.

Nemec MD, Kaufman LE, Stump DG, Lindstrom P, Varsho BJ, Holson JF. 2006. Significance, Reliability, and Interpretation of Developmental and Reproductive Toxicity Study Findings. In: Developmental Reproductive Toxicology: A Practical Approach:Informa Healthcare.

Newman LM, Johnson EM, Staples RE. 1993. Assessment of the Effectiveness of Animal Developmental Toxicity Testing for Human Safety. Reprod Toxicol 7(4): 359-390.

Nisbet ICT, Karch NJ. 1983. Chemical Hazards to Human Reproduction. Park Ridge, NJ: Noyes Data Corp.

Peterson ED. 2008. Research methods to speed the development of better evidence- the registries example. In: Evidence-based medicine and the changing nature of health care:2007 IOM annual meeting summary. Washington, DC:The National Academies Press.

Viswanathan M, Ansari M, Berkman N, Chang S, Hartling L, McPheeters L, et al. 2012. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. (Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews). AHRQ Publication No. 12-EHC047-EF.